

ADVENTURE IN AMERICAN EDUCATION

 **VOLUME THREE** 

SMITH AND TYLER

17-8

DTN 7
881/

37125
6 M I

~~371.29~~
~~5644~~



PROGRESSIVE EDUCATION ASSOCIATION PUBLICATIONS
Commission on the Relation of School and College

ADVENTURE
IN
AMERICAN EDUCATION

Volume III
Appraising and Recording Student Progress

Bureau Ednl. & Res. Research

DAVID HALL - TRAINING COLLEGE

Dated.....3.9.55.....

Accs. No...839.....

The Commission on the Relation of School and College
of
The Progressive Education Association

MEMBERS OF THE COMMISSION

Walter Raymond Agard	John A. Lester
Wilford M. Aikin, Chairman	Max McConn, Secretary
Willard Beatty	Clyde R. Miller
Bruce Bliven	°Jesse H. Newlon
C. S. Boucher	W. Carson Ryan
°A. J. Burton	Harold Rugg
Flora S. Cooke	°Ann Shumaker
Harold A. Ferguson	Eugene R. Smith
Burton P. Fowler	Perry Dunlap Smith
Josephine Gleason	Katharine Taylor
Thomas Hopkins	Vivian T. Thayer
Leonard V. Koos	Goodwin Watson
W. S. Learned	Raymond Walters
Robert D. Leigh	Ben D. Wood

After originating and organizing the Eight-Year Study, the Commission in 1933 gave full responsibility and authority for the supervision of the Study to the Directing Committee.

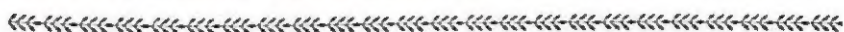
DIRECTING COMMITTEE OF THE COMMISSION

Wilford M. Aikin, Chairman, Director of the Study	Robert D. Leigh
Willard Beatty	Max McConn, Secretary
Boyd H. Bode	°Jesse H. Newlon
Burton P. Fowler	Marion Park
Carl Brigham	Eugene R. Smith
Will French	J. E. Stonecipher
Herbert E. Hawkes	¹ John B. Johnson
John A. Lester	
Elizabeth M. Steel, Secretary to the Director	

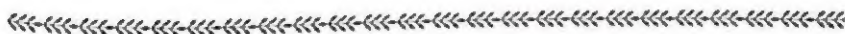
° Deceased.

¹ Resigned.

ADVENTURE IN AMERICAN EDUCATION VOLUME III



APPRAISING AND RECORDING STUDENT PROGRESS



EUGENE R. SMITH, RALPH W. TYLER
and the EVALUATION STAFF

McGRAW-HILL BOOK COMPANY, INC.
New York Toronto London

APPRAISING AND RECORDING STUDENT PROGRESS

Copyright, 1942, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

371.27

372

The Progressive Education Association
the Commission and the Schools
gratefully acknowledge their indebtedness
to

CARNEGIE CORPORATION OF NEW YORK

and to the

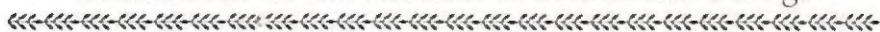
GENERAL EDUCATION BOARD

for the funds which made this Study possible

During the first year the Commission had no funds except \$800 contributed in equal amounts by the Francis W. Parker, John Burroughs, Lincoln, and Tower Hill Schools. From the beginning of 1932, generous subventions from Carnegie Corporation of New York supported the work, except that in evaluation, through 1936. Much larger grants from the General Education Board financed the work of the Evaluation Staff, the Curriculum Associates and, since 1936, all the activities of the Commission.

Neither Carnegie Corporation of New York nor the General Education Board is author, owner, publisher, or proprietor of this publication. They are not to be understood as approving by virtue of their grants any of the statements made or views expressed therein.

Commission on the Relation of School and College



COMMITTEE ON EVALUATION AND RECORDING

Helen M. Atkinson	*Frances Knapp
Frederick H. Bair	Robert D. Leigh
E. Gordon Bill	Max McConnell
Burton P. Fowler	Eugene R. Smith, Chairman
Ben D. Wood	

SUB-COMMITTEES¹ ON RECORDS AND REPORTS

Committee on Behavior Description

Helen M. Atkinson	Anna Rose Hawkes	Rollo G. Reynolds
E. Gordon Bill	*Frances Knapp	Eugene R. Smith
Carl Brigham	Robert D. Leigh	John Tildsley
Oscar K. Buros	W. S. Learned	Ben D. Wood
Cecile Flemming	John A. Lester	Stanley R. Yarnall
	John W. M. Rothney	

Committee on Teachers' Reports and Reports to the Home

Helen M. Atkinson	Rosamond Cross	Elmina R. Lucke
Derwood Baker	Burton P. Fowler	Eugene R. Smith
Genevieve L. Coy	I. R. Kraybill	John W. M. Rothney

Committee on Form for Transfer from School to College

Victor L. Butterfield	Ruth W. Crawford	Eugene R. Smith
Genevieve L. Coy	Burton P. Fowler	Herbert W. Smith
Albert B. Crawford	Elmina R. Lucke	Arthur E. Traxler
	John W. M. Rothney	

General Committee on Study of the Development of Pupils in Subject Fields

Helen M. Atkinson	Edith M. Penney
Genevieve L. Coy	Eugene R. Smith
Harry Herron	Arthur E. Traxler
G. H. V. Melone	John W. M. Rothney

* Deceased.

¹ In addition to those who were continuing committee members, at least 400 others from the schools and other institutions cooperated.

Commission on the Relation of School and College

EVALUATION STAFF

Ralph W. Tyler, Research Director

Associate Director

Oscar K. Buros, 1934-35
Louis E. Raths, 1935-38
Maurice L. Hartung, 1938-42

Associates

Bruno Bettelheim
Paul B. Diederich
Wilfred Eberhart
Louis M. Heil
George Sheviakov
Hilda Taba
Harold Trimble

Assistants

Herbert J. Abraham
Dwight L. Arnold
Jean Friedberg Block
Charles L. Boye
Fred P. Frutchey
Paul R. Grim
Chester William Harris
John H. Herrick
Clark W. Horton
Walter Howe
Carleton C. Jones
W. Harold Lauritsen
Christine McGuire
Harold G. McMullen
Donald H. McNassor
Secretaries: Cecelia K. Wasserstrom, Kay D. Watson

CURRICULUM ASSOCIATES

H. H. Giles

S. P. McCutchen

A. N. Zechiel

The following served as special curriculum consultants at various times:

Harold B. Alberty
Paul B. Diederich
John A. Lester
Henry Harap
Walter V. Kaulfers

COLLEGE FOLLOW-UP STAFF

John L. Bergstresser
Dean Chamberlin
Enid Straw Chamberlin
Neal E. Drought
William E. Scott
Harold Threlkeld

EDITORIAL COMMITTEE

Harold B. Alberty
C. L. Cushman
Burton P. Fowler
Max McConn
Thomas C. Pollock

Commission on the Relation of School and College

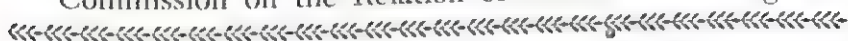
THE PARTICIPATING SCHOOLS

<i>School</i>	<i>Head¹</i>
Altoona Senior High School, Altoona, Pa.	(Levi Gilbert) Joseph N. Maddocks
Baldwin School, Bryn Mawr, Pa.	(Miss Elizabeth Johnson) Miss Ros- amond Cross
Beaver Country Day School, Chest- nut Hill, Mass.	Eugene R. Smith
Bronxville High School, Bronxville, N. Y.	Miss Edith M. Penney
Cheltenham Township High School, Elkins Park, Pa.	I. R. Kraybill
Dalton Schools, New York, N. Y.	Miss Helen Parkhurst
Denver Senior and Junior High Schools, Denver, Colo.	(Charles Greene) (C. L. Cushman) John J. Cory.
Des Moines Senior and Junior Schools, Des Moines, Iowa	(*R. C. Cook) J. E. Stonecipher
Eagle Rock High School, Los An- geles, Cal.	Miss Helen Babson
Fieldston School, New York, N. Y.	(Herbert W. Smith) (Derwood Baker) Luther Tate
Francis W. Parker School, Chicago, Ill.	(Miss Flora Cooke) (Raymond Os- borne) Herbert W. Smith
Friends' Central School, Overbrook, Pa.	Barclay L. Jones
George School, George School, Pa.	George A. Walton
Germantown Friends School, Ger- mantown, Pa.	(Stanley R. Yarnall) Burton P. Fowler
Horace Mann School, New York, N. Y.	(Rollo G. Reynolds) Will French
John Burroughs School, Clayton, Mo.	(Wilford M. Aikin) Leonard D. Haertter
Lincoln School of Teachers College, New York, N. Y.	(*Jesse H. Newlon) (Lester Dix) Will French
Milton Academy, Milton, Mass.	W. L. W. Field
New Trier Township High School, Winnetka, Ill.	Matthew P. Gaffney
North Shore Country Day School, Winnetka, Ill.	Perry Dunlap Smith
Radnor High School, Wayne, Pa.	Sydney V. Rowland, T. Bayard Beatty

¹ Many changes in administration occurred in the schools during the period of the Study. Such cases are indicated by names in parentheses given in chronological order of service.

* Deceased

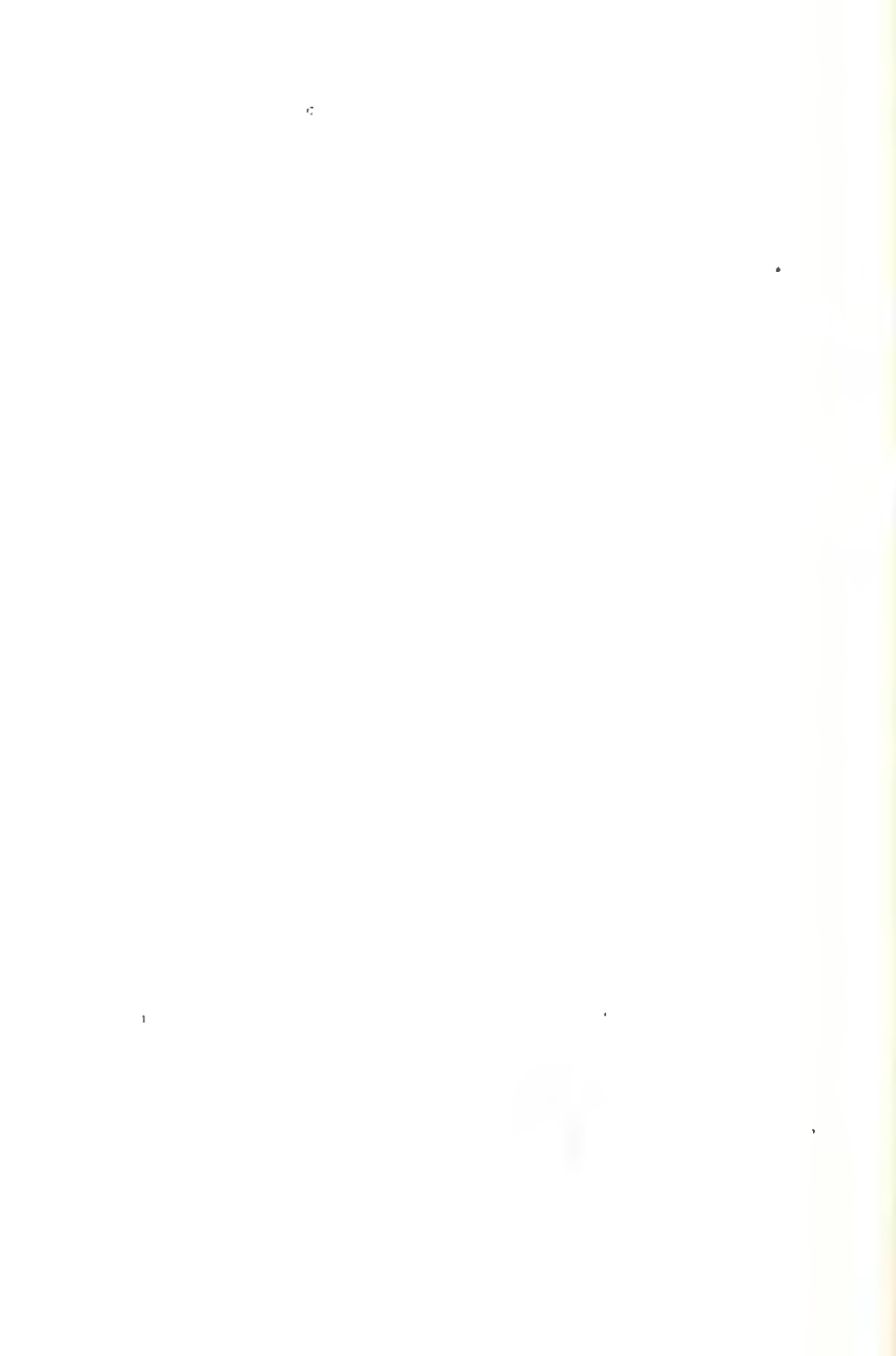
Commission on the Relation of School and College



School

Head

Shaker High School, Shaker Heights, Ohio	R. B. Patin
Tower Hill School, Wilmington, Del.	(Burton P. Fowler) James S. Guernsey
Tulsa • Senior and Junior High Schools, Tulsa, Okla.	(Will French) (Eli C. Foster) H. W. Cowans
University of Chicago High School, Chicago, Ill.	(Arthur K. Loomis) P. B. Jacobson
University High School, Oakland, Cal.	(George Rice) Paul T. Fleming
University School of Ohio State Uni- versity, Columbus, Ohio	(Rudolph Lindquist) (Harold B. Alberty) Robert S. Gilchrist
Winsor School, Boston, Mass.	(Miss Katharine Lord) Miss Frances Dugan
Wisconsin High School, Madison, Wisc.	(H. H. Ryan) (Stephen M. Corey) (Gordon Mackenzie) Glen G. Eye



CONTENTS

FOREWORD	xvii
PREFACE	xxi

Part I—Development and Use of Evaluation Instruments

I. PURPOSES AND PROCEDURES OF THE EVALUATION STAFF, <i>Ralph W. Tyler</i>	3
How the Evaluation Staff Came into Existence	3
Significance of the Evaluation Project	5
Major Purposes of Evaluation	7
Basic Assumptions	11
General Procedures in Developing the Evaluation Program	15
Division of Labor in the Evaluation Program	28
Summary	29
II. ASPECTS OF THINKING, <i>Maurice L. Hartung, Leah Weisman, Harold G. McMullen, Harold C. Trimble</i>	35
I. Interpretation of Data—Analysis of the Objec- tive	38
The Development of Evaluation Instruments	43
Other Instruments to Measure This Objective	60
Validity of the Interpretation of Data Tests	65
Reliability of the Interpretation of Data Tests	74
II. Application of Principles of Science—Analysis ^o of the Objective	77
The Development of Evaluation Instruments	80
III. Application of Principles of Logical Reason- ing—Analysis of the Objective	111
The Development of Evaluation Instruments	114
Validity and Reliability of Form 5.12	124
IV. The Nature of Proof—Analysis of the Objec- tive	126

The Development of Evaluation Instruments	130
Validity and Reliability of Test Form 5.22	143
A Related Instrument	148
Conclusion	154
III. EVALUATION OF SOCIAL SENSITIVITY, <i>Hilda Taba,</i> <i>Christine McGuire</i>	
Informal Methods of Getting Evidence on Social Sensitivity	157
Evaluation of the Ability to Apply Social Facts and Generalizations	162
Analysis of the Objective	168
Construction of the Test on the Ability to Apply Social Values	175
Validity and Reliability	190
Applying Social Facts and Generalizations to Social Problems (Form 1.5)	197
Evaluation of Social Attitudes—Analysis of the Objective	203
Evaluation of Beliefs on Social Issues	209
Description of the Test on Beliefs on Social Issues (Form 4.21-4.31)	215
Beliefs about School Life	229
Beliefs on Economic Issues	234
Uses of These Tests	239
IV. ASPECTS OF APPRECIATION, <i>Chester William Harris,</i> <i>Bruno Bettelheim, Paul B. Diederich</i>	
Appreciation of Literature	245
The Evaluation of the Appreciation of Art	246
Development of the Instrument	276
Description of the Test	283
The Test Interpretation	289
How to Administer the Test	292
Reliability and Validity	299
Future Use of the Test	300
Other Instruments	306
V. EVALUATION OF INTERESTS, <i>Chester William Harris,</i> <i>Wilfred Eberhart, Jean Friedberg Block</i>	
	313

CONTENTS

xv

Analysis of the Objective	313
The Interest Index 8.2a	338
VI. EVALUATION OF PERSONAL AND SOCIAL ADJUSTMENT, <i>George V. Sheviakov, Jean Friedberg Block</i>	349
Discussion of the Objective	349
Discussion of the Technique of Appraisal of the Objective	354
Description of the Questionnaire	361
Interpretation of the Responses to the Question- naires	372
Sample Analysis of Response of One Student	377
Reliability	384
Validity	385
Possible Uses of the Questionnaires	396
Summary	400
VII. INTERPRETATION AND USES OF EVALUATION DATA, <i>Hilda Taba</i>	403
Methods of Interpreting and Using Evaluation Data	430
VIII. PLANNING AND ADMINISTERING THE EVALUATION PRO- GRAM, <i>Hilda Taba</i>	441
Part II—Recording for Guidance and Transfer	
<i>Eugene R. Smith</i>	
IX. PHILOSOPHY AND OBJECTIVES	463
General Purposes and Philosophy of Recording	464
Working Objectives for Records and Reports	467
X. BEHAVIOR DESCRIPTION	470
Use of Record Cards	474
Summary of Advantages	486
XI. TEACHERS' REPORTS AND REPORTS TO THE HOME	488
XII. FORM FOR TRANSFER FROM SCHOOL TO COLLEGE	494
Confidential Report to the Committee on Admission	494
The "Junior Year" Blank	498
XIII. STUDY OF THE DEVELOPMENT OF PUPILS IN SUBJECT FIELDS	499
APPENDICES	507

FOREWORD

It is an amazing fact that our schools and colleges know little of the results of their work. It is even more amazing that they seldom attempt seriously to find out what changes schooling brings about in students. Ask any school what its objectives are and you will be told that it seeks to develop character, ability to think clearly, social responsibility, good health habits, readiness for earning a living, knowledge of certain facts and mastery of certain skills. Ask whether the school succeeds in doing these things, the answer is, "We know only in part." Half of the boys and girls who begin the work of the secondary school drop out before completing it. Schools usually do not know why these students leave or what becomes of them immediately afterward. Few schools know even what their graduates are doing, what problems they are facing, or how well prepared they are to solve them.

How can this lack of knowledge and concern be explained? There are doubtless many causes, but one of the most obvious is the universal emphasis upon the accumulation of credits for promotion, graduation, and admission to college. To secure a credit or unit the student must "pass" a course. To pass a course he must remember certain facts and show proficiency in certain skills. Therefore, remembering knowledge and practicing techniques for examinations become the purposes of education for pupils and teachers alike. What goes on the school record becomes the real objective of the student, no matter what the school *says* its purposes are. If the pupil secures the required credits, he is graduated. The job is done. Concentration on these worthy but limited goals seems to make teachers and students forget the larger, long-range purposes of education.

One of the major reasons for over-emphasis upon these limited objectives is that results in these fields are more easily measured than in other less tangible areas. There are many instruments of evaluation applicable to the conventional subjects of the curriculum. Much of the work of such organizations as the Educational Records Bureau, the Co-operative Test Service, and the College Entrance Examination Board is of great value to schools and colleges. But most tests available when this Study began were measures chiefly of accretions of knowledge and proficiency in the use of skills. Because such tests are at hand the teacher uses them. Because instruments of appraisal in other areas have not been available, the teacher tends to neglect other objectives and to strive only for results that can be ascertained with relative ease and objectivity.

It follows, then, that comprehensive appraising, recording, and reporting of results are matters of vital concern to those who seek improvement in the work of our schools and colleges. The Eight-Year Study has recognized the importance of these aspects of school work. To assist the Thirty Schools in developing adequate programs of evaluation and reporting, committees and technical staffs were organized shortly after the Study began. The Commission was fortunate in securing the services of Eugene R. Smith and Ralph W. Tyler as leaders in this work. This volume reports in detail the steps that were taken to help the schools to discover, record, and report the progress of students toward the whole range of desired goals.

The work reported here rests upon three basic convictions: first, that evaluation and recording should always be directly related to each school's purposes; second, that any school's evaluation program should be comprehensive, including appraisal of progress toward *all* the school's major objectives; third, that teachers should participate in the construction of all instruments of evaluation and forms for records and reports.

It is impossible to estimate the wastage of material and human resources which results from education's ignorance of the consequences of its efforts. Until schools and colleges develop adequate, comprehensive appraising and recording programs, that waste will continue. Although no one connected with the Eight-Year Study would claim that its work in these fields is complete or entirely satisfactory, it is clear that what is reported in this volume points the way to fuller knowledge, more complete understanding, and wiser guidance of youth.

WILFORD M. AIKIN

PREFACE

When the Directing Committee of the Commission on the Relation of School and College appointed a Committee on Records and Reports, it assigned to this new committee the general task of recommending methods of obtaining and recording information about the pupils. The immediate reason for this assignment was the need of supplying to the colleges data upon which they could decide about the acceptability of candidates who did not present the traditional pattern of subjects for entrance or had not submitted the usual entrance information in terms of marks and examinations. A second important reason was the desire of schools for help in their guidance programs.

The instructions given this committee specified as its first task the devising of methods of obtaining and recording information about personality. It was necessary, however, from the beginning to try to find ways of testing that would neither determine nor depend upon the content of the courses given in the various schools, yet would be reasonably comparable and objective measures of knowledge and power.

The committee met with some frequency for periods of two or three days at a time. It soon announced to the schools a list of comparable tests that seemed to have value for estimating the degree of mastery attained by pupils in various subject fields. Many of the schools tried these tests, and some added others from quite a wide selection of those of an objective type. It became apparent, however, that even these tests were too much influenced by the content studied to be acceptable to all of the schools. The reason was that the schools were anxious to use the utmost flex-

ibility in meeting the needs of their pupils even when that meant departing markedly from traditional subjects or their content. A period of experimentation followed, during which other work was accomplished. When it was recognized that no matter how valuable existing methods and material for testing might be for various purposes, nevertheless they did not fit the need of the cooperating schools for testing that would measure the *power attained, irrespective of the way in which it had been reached*, the Directing Committee obtained further funds and enlarged the branch responsible for testing, recording, and reporting.

The final organization of this department was headed by an over-all committee called the Committee on Evaluation and Recording. It had responsibility for determining policies, considering reports on work accomplished and giving direction about the next steps to be undertaken. Dr. Ralph W. Tyler was engaged as Research Director for this part of the Eight-Year Study, and was given as his particular assignment charge of the work on evaluation. This assignment included direction of the follow-up study of graduates of the cooperating schools who were attending college, as well as of the study of objectives and of the testing and other evaluation carried on in the schools. Under Dr. Tyler's supervision the Evaluation Staff and a large number of committees assisted in this part of the work. A detailed account is given in Part I of this volume.

The chairman of the Committee on Evaluation and Recording was given charge of the production of recording forms, and of methods of reporting to the colleges and to the homes. As a part of this work the original Committee on Records and Reports, which had in the meantime published two editions of the "Behavior Description," described in Part II, was assigned the continued study of personal characteristics and their recording and reporting.

Other committees whose members were chosen not only

from the cooperating schools but also from colleges and from schools and other groups not definitely concerned with the Eight-Year Study, worked on the various problems concerned with records and reports and were responsible for the forms devised. Of much importance also was the help given by the various members of the staff. The assistance of the Director of the Study, the Research Director, the Curriculum Assistants and the Members of the Evaluation Staff was available both indirectly, through the results of their studies, and directly by means of conferences and attendance at group meetings. Dr. John W. M. Rothney deserves special mention since he has been Research Assistant to all of these committees since the change in organization.

While it is not possible to list the large number of those who took part in the work on evaluation and that on recording, the committee in charge of these activities wishes to express its appreciation of the contributions made by those who assisted. Without their self-sacrificing cooperation, little could have been accomplished.

EUGENE R. SMITH,
Chairman

PART I

DEVELOPMENT AND USE OF EVALUATION
INSTRUMENTS



Chapter I

PURPOSES AND PROCEDURES OF THE EVALUATION STAFF

HOW THE EVALUATION STAFF CAME INTO EXISTENCE

The plan of the Eight-Year Study, as Dr. Smith explained in the Preface, placed upon the cooperating schools the responsibility for reporting in some detail the characteristics and achievements of students who were recommended for admission to college. Furthermore, the Directing Committee of the Study expected the schools not only to record the steps taken to develop new educational programs, but also to appraise the effectiveness of these programs, so that other schools might benefit from their experience.

After the first year it became clear that these tasks were too great for them both to be assumed by the Committee on Records and Reports. The magnitude of the work had become evident when the Committee on Records and Reports reviewed the available tests, examinations, and other devices for appraising student achievement. Most of the achievement tests then on the market measured only the amount of information which students remembered, or some of the more specific subject skills like those in algebra and the foreign languages. The new courses developed in the Thirty Schools attempted to help students achieve several additional qualities, such as more effective study skills, more careful ways of thinking, a wider range of significant interests, social rather than selfish attitudes. Hence, the available achievement tests did not provide measures of many of the more important achievements anticipated from these new

courses. Furthermore, the content of most significance in the new courses was frequently different from that which had been included before. Hence, the available tests of information did not really measure the information which students would be obtaining in the new courses. A comprehensive appraisal of the new educational programs could not be carried on unless new means of evaluating achievement were developed.

The Directing Committee obtained a preliminary subsidy from the General Education Board to explore the possibility of constructing devices which could be used in appraising the outcomes of the new work. During the autumn of 1934, the Thirty Schools were visited, inter-school committees were formed, and preliminary steps taken to construct needed instruments of evaluation. By the winter of 1935 it seemed apparent that new instruments could be devised and that a more comprehensive program of appraisal could be conducted. Hence, a generous subsidy for the services of an evaluation staff¹ was provided by the General Education Board, and the work was continued until the close of the

¹ During the exploratory period, Oscar K. Buros, of Rutgers University, served as Associate Director. After helping to get the plan outlined, Mr. Buros resigned as Associate Director of the Evaluation Staff and returned to Rutgers University. From July, 1935, until September, 1938, Mr. Louis E. Rathis served as Associate Director. The Staff was then housed at the Ohio State University. When Mr. Tyler, the Director, moved to the University of Chicago in September, 1938, Mr. Maurice L. Hartung was made Associate Director. Others who served as members of the staff at least part time for one or more years were: Herbert J. Abraham, Dwight L. Arnold, Bruno Bettelheim, Jean Friedberg Block, Charles L. Boye, Paul B. Diederich, Wilfred Eberhart, Fred P. Frutchey, Paul R. Grim, Chester William Harris, Louis M. Heil, John H. Herrick, Clark W. Horton, Walter Howe, Carleton C. Jones, W. Harold Lauritsen, Christine McGuire, Harold G. McMullen, Donald H. McNassor, George V. Sheviakov, Hilda Taba, Harold Trimble, Cecelia K. Wasserstrom, Kay D. Watson, Leah Weisman.

Throughout the years these persons have worked together as a unified staff. Although authorship of chapters is indicated in the table of contents, in a very real sense this report is a staff document, the product of all members of the staff. Each chapter was criticized and revised several times by all those who were members of the staff at the time the report was written.

Study. The Evaluation Staff was primarily^o concerned with developing means by which the achievement of the students in the schools could be appraised, and the strengths and weaknesses of the school programs could be identified.

In 1936 the first class enrolled in these new programs graduated from the Thirty Schools, and most of them entered college in the fall. This provided an opportunity to appraise the school programs in terms of the success of their graduates in college. Through the generosity of the General Education Board, funds were provided for this study and a second division of the Evaluation Staff² was established. The report of the study of college success appears in another volume. The present volume is devoted to the discussion of evaluation in the Schools and methods of recording and reporting.

SIGNIFICANCE OF THE EVALUATION PROJECT

The term "evaluation" was used to describe the staff and the project rather than the term "measurement," "test," or "examination" because the term "evaluation" implies a process by which the values of an enterprise are ascertained. To help provide means by which the Thirty Schools could ascertain the values of their new programs was the basic purpose of the evaluation project. The project has significance not only for the Thirty Schools but for schools and colleges generally. Adequate appraisal of the educational program of a school or college is rarely made. Yet an appraisal of an educational institution is fundamentally only the process^o by which we find out how far the objectives of the institution are being realized. This seems a simple and straightforward task, and the efforts at evaluation of certain social institutions are not very complex. For example, in the case of a retail business enterprise the most commonly recognized ob-

² Composed of John L. Bergstresser, Dean Chamberlin, Enid Straw Chamberlin, Neal Drought, William E. Scott, Harold Threlkeld.

jectives are two: namely, the distribution of large quantities of goods and the making of profit from the sale of these goods. The methods for determining the quantities of goods sold and the profits are tangible and not very difficult to apply. Hence, the problem of evaluation is not usually considered a perplexing one, and although the business enterprise devotes a portion of its time and energy to appropriate accounting procedures, so as to make a periodical evaluation of its activities, we do not find a high degree of uncertainty about the methods of evaluation.

In education, however, the problem of evaluation is more complex for several reasons. In the first place, since schools generally have not agreed upon their fundamental objectives, there is doubt as to what values schools expect to attain and therefore what results to look for in the process of evaluation. Even when the objectives of a school are agreed upon and stated, they are frequently vague and require clarification in order to be understood. Furthermore, the methods of obtaining evidence about the attainment of some of these educational objectives are more difficult and less direct processes than those used in appraising a business. It is easy to see how to measure the amount of profit in a retail store; it is not so easy to devise ways for measuring the educational changes taking place in students in the school. Finally, the task of summarizing and interpreting the results of an evaluation of the school is complicated. Summaries of educational evaluation are needed for several different groups, that is, for students, teachers, administrators, parents, and patrons. Each of these groups may need somewhat different information, or at least it will be necessary to present the data in different terms. It is easy to see, then, that educational evaluation requires more intensive study than evaluation of many other institutions. The work of the Evaluation Staff should help to demonstrate procedures by which the process of evaluation may be carried on and to

provide instruments and devices that may be used in evaluation or that may suggest ideas for the construction of other instruments.

MAJOR PURPOSES OF EVALUATION

In perceiving the appropriate place of evaluation in modern education, consideration must be given to the purposes which a program of evaluation may serve. At present the purposes most commonly emphasized in schools and colleges are the grading of students, their grouping and promotion, reports to parents, and financial reports to the board of education or to the board of trustees. A comprehensive program of evaluation should serve a broader range of purposes than these.

One important purpose of evaluation is to make a periodic check on the effectiveness of the educational institution, and thus to indicate the points at which improvements in the program are necessary. In a business enterprise the monthly balance sheet serves to identify those departments in which profits have been low and those products which have not sold well. This serves as a stimulus to a re-examination and a revision of practices in the retail establishment. In a similar fashion, a periodic evaluation of the school or college, if comprehensively undertaken, should reveal points of strength which ought to be continued and points where practices need modification. This is helpful to all schools, not just to schools which are experimenting.

A very important purpose of evaluation which is frequently not recognized is to validate the hypotheses upon which the educational institution operates. A school, whether called "traditional" or "progressive," organizes its curriculum on the basis of a plan which seems to the staff to be satisfactory, but in reality not enough is yet known about curriculum construction to be sure that a given plan will work satisfactorily in a particular community. On that ac-

count, the curriculum of every school is based upon hypotheses, that is, the best judgments the staff can make on the basis of available information. In some cases these hypotheses are not valid, and the educational institution may continue for years utilizing a poorly organized curriculum because no careful evaluation has been made to check the validity of its hypotheses. For example, many high schools and colleges have constructed the curriculum on the hypothesis that students would develop writing habits and skills appropriate to all their needs if this responsibility were left entirely to the English classes. Careful appraisal has shown that this hypothesis is rarely, if ever, valid. Similarly, in a program of guidance the effort to care for personal and social maladjustments among students in a large school is sometimes based on the hypothesis that the provision of a well-trained guidance officer for the school will eliminate maladjustments. Systematic evaluation has generally shown that one officer has little effect unless a great deal of supplementary effort is devoted to educating teachers in child development and to revising the curriculum at those points where it promotes maladjustments. In the same way, many of our administrative policies and practices are based upon judgments which in a particular case may not be sound. Every educational institution has the responsibility of testing the major hypotheses upon which it operates and of adding to the fund of tested principles upon which schools may better operate in the future.

A third important purpose of evaluation is to provide information basic to effective guidance of individual students. Only as we appraise the student's achievement and as we get a comprehensive description of his growth and development are we in a position to give him sound guidance. This implies evaluation sufficiently comprehensive to appraise all significant aspects of the student's accomplishments. Merely the judgment that he is doing average work in a

particular course is not enough. We need to find out more accurately where he is progressing and where he is having difficulties.

A fourth purpose of evaluation is to provide a certain psychological security to the school staff, to the students, and to the parents. The responsibilities of an educational institution are broad and involve aspects which seem quite intangible to the casual observer. Frequently the staff becomes a bit worried and is in doubt as to whether it is really accomplishing its major objectives. This uncertainty may be a good thing if it leads to a careful appraisal and constructive measures for improvement of the program; but without systematic evaluation the tendency is for the staff to become less secure and sometimes to retreat to activities which give tangible results although they may be less important. Often we seek security through emphasizing procedures which are extraneous and sometimes harmful to the best educational work of the school. Thus, high school teachers may devote an undue amount of energy to coaching for scholarship tests or college entrance examinations because the success of students on these examinations serves as a tangible evidence that something has been accomplished. However, since these examinations may be appropriate for only a portion of the high school student body, concentration of attention upon them may actually hinder the total educational program of the high school. For such teachers a comprehensive evaluation which gives a careful check on all aspects of the program would provide the kind of security that is necessary for their continued growth and self-confidence. This need is particularly acute in the case of teachers who are developing and conducting a new educational program. The uncertainty of their pioneering efforts breeds insecurity. They view with dismay or resentment efforts to appraise their work in terms of devices appropriate only to the older, previously established curriculum. They

recognize that the effectiveness of the new work can be fairly appraised only in terms of its objectives, which in certain respects differ from the purposes of the older program. Students and parents are also subject to this feeling of insecurity and in many cases desire some kind of tangible evidence that the educational program is effective. If this is not provided by a comprehensive plan of evaluation, then students and parents are likely to turn to tangible but extraneous factors for their security.

A fifth purpose of evaluation which should be emphasized is to provide a sound basis for public relations. No factor is as important in establishing constructive and cooperative relations with the community as an understanding on the part of the community of the effectiveness of its educational institutions. A careful and comprehensive evaluation should provide evidence that can be widely publicized and used to inform the community about the value of the school or college program. Many of the criticisms expressed by patrons and parents can be met and turned to constructive cooperation if concrete evidence is available regarding the accomplishments of the school.

Evaluation can contribute to these five purposes. It can provide a periodic check which gives direction to the continued improvement of the program of the school; it can help to validate some of the important hypotheses upon which the program operates; it can furnish data about individual students essential to wise guidance; it can give a more satisfactory foundation for the psychological security of the staff, of parents, and of students; and it can supply a sound basis for public relations. These purposes were basic to the Thirty Schools but they are also important to all schools. For these purposes to be achieved, however, they must be kept continually in mind in planning and in developing the program of evaluation. The Evaluation Staff realized that the decision as to what is to be evaluated, the

techniques for appraisal, and the summary⁹ and interpretation of results should all be worked out in terms of these important purposes.

BASIC ASSUMPTIONS

In developing the program, the Evaluation Staff accepted certain basic assumptions. Eight of them were of particular importance. In the first place, it was assumed that education is a process which seeks to change the behavior patterns of human beings. It is obvious that we expect students to change in some respects as they go through an educational program. An educated man is different from one who has no education, and presumably this difference is due to the educational experience. It is also generally recognized that these changes brought about by education are modifications in the ways in which the educated man reacts, that is, changes in his ways of behaving. Generally, as a result of education we expect students to recall and to use ideas which they did not have before, to develop various skills, as in reading and writing, which they did not previously possess, to improve their ways of thinking, to modify their reactions to esthetic experiences as in the arts, and so on. It seems safe to say on the basis of our present conception of learning, that education, when it is effective, changes the behavior patterns of human beings.

A second basic assumption was that the kinds of changes in behavior patterns in human beings which the school seeks to bring about are its educational objectives. The fundamental purpose of an education is to effect changes in the behavior of the student, that is, in the way he thinks, and feels, and acts. The aims of any educational program cannot well be stated in terms of the content of the program or in terms of the methods and procedures followed by the teachers, for these are only means to other ends. Basically, the goals of education represent these changes in human

beings which we hope to bring about through education. The kinds of ideas which we expect students to get and to use, the kinds of skills which we hope they will develop, the techniques of thinking which we hope they will acquire, the ways in which we hope they will learn to react to esthetic experiences—these are illustrations of educational objectives.

A third basic assumption was referred to at the opening of the chapter. An educational program is appraised by finding out how far the objectives of the program are actually being realized. Since the program seeks to bring about certain changes in the behavior of students, and since these are the fundamental educational objectives, then it follows that an evaluation of the educational program is a process for finding out to what degree these changes in the students are actually taking place.

The fourth basic assumption was that human behavior is ordinarily so complex that it cannot be adequately described or measured by a single term or a single dimension. Several aspects or dimensions are usually necessary to describe or measure a particular phase of human behavior. Hence, we did not conceive that a single score, a single category, or a single grade would serve to summarize the evaluation of any phase of the student's achievement. Rather, it was anticipated that multiple scores, categories, or descriptions would need to be developed.

The fifth assumption was a companion to the fourth. It was assumed that the way in which the student organizes his behavior patterns is an important aspect to be appraised. There is always the danger that the identification of these various types of objectives will result in their treatment as isolated bits of behavior. Thus, the recognition that an educational program seeks to change the student's information, skills, ways of thinking, attitudes, and interests, may result in an evaluation program which appraises the development

of each of these aspects of behavior separately, and makes no effort to relate them. We must not forget that the human being reacts in a fairly unified fashion; hence, in any given situation information is not usually separated from skills, from ways of thinking, or from attitudes, interests, and appreciations. For example, a student who encounters an important social-civic problem is expected to draw upon his information, to use such skill as he has in locating additional facts, to think through the problem critically, to make choices of courses of action in terms of fundamental values and attitudes, and to be continually interested in better solutions to such problems. This clearly involves the relationship of various behavior patterns and their better integration. The way the student grows in his ability to relate his various reactions is an important aspect of his development and an important part of any evaluation of his educational achievement.

A sixth basic assumption was that the methods of evaluation are not limited to the giving of paper and pencil tests; any device which provides valid evidence regarding the progress of students toward educational objectives is appropriate. As a matter of practice, most programs of appraisal have been limited to written examinations or paper and pencil tests of some type. Perhaps this has been due to the long tradition associated with written examinations or perhaps to the greater ease with which written examinations may be given and the results summarized. However, a consideration of the kinds of objectives formulated for general education makes clear that written examinations are not likely to provide an adequate appraisal for all of these objectives. A written test may be a valid measure of information recalled and ideas remembered. In many cases, too, the student's skill in writing and in mathematics may be shown by written tests, and it is also true that various techniques of thinking may be evidenced through more novel types of

written test materials. On the other hand, evidence regarding the improvement of health practices, personal-social adjustment, interests, and attitudes may require a much wider repertoire of appraisal techniques. This assumption emphasizes the wider range of techniques which may be used in evaluation, such as observational records, anecdotal records, questionnaires, interviews, check lists, records of activities, products made, and the like. The selection of evaluation techniques should be made in terms of the appropriateness of these techniques for the kind of behavior to be appraised.

A seventh basic assumption was that the nature of the appraisal influences teaching and learning. If students are periodically examined on certain content, the tendency will be for them to concentrate their study on this material, even though this content is given little or no emphasis in the course of study. Teachers, too, are frequently influenced by their conception of the achievement tests used. If these tests are thought to emphasize certain points, these points will be emphasized in teaching even though they are not included in the plan of the course. This influence of appraisal upon teaching and learning led the Evaluation Staff to try to develop evaluation instruments and methods in harmony with the new curricula and, as far as possible, of a non-restrictive nature. That is, major attention was given to appraisal devices appropriate to a wide range of curriculum content and to varied organizations of courses. Much less effort was devoted to the development of subject-matter tests since these assumed certain common informational material in the curriculum.

The eighth basic assumption was that the responsibility for evaluating the school program belonged to the staff and clientele of the school. It was not the duty of the Evaluation Staff to appraise the school but rather to help develop the means of appraisal and the methods of interpretation. Hence, this volume does not contain an appraisal of the work

of the Thirty Schools or the results obtained by the use of the evaluation instruments in the schools. This volume is a report of the development of techniques for evaluation.

The evaluation program utilized other assumptions but these eight were of particular importance because they guided the general procedure by which the evaluation program was developed. They showed the necessity for basing an evaluation program upon educational objectives, and they indicated that educational objectives for purposes of evaluation must be stated in terms of changes in behavior of students; they emphasized the multiple aspects of behavior and the importance of the relation of these various aspects of behavior rather than treatment of them in isolation; and they made clear the possibility of a wide range of evaluation techniques.

GENERAL PROCEDURES IN DEVELOPING THE EVALUATION PROGRAM

The general procedure followed in developing the evaluation program involved seven major steps. Since the program was a cooperative one, including both the Schools and the Evaluation Staff, it should be clear that although the report was prepared by the staff, the work was done by a large number of persons. No one of the instruments developed is the product of a single author. All have required the efforts of various members of the school staffs and the Evaluation Staff.

1. *Formulating Objectives*

As the first step, each school faculty was asked to formulate a statement of its educational objectives. Since the schools were in the process of curriculum revision, several of them had already taken this step. This is not just an evaluation activity; for it is usually considered one of the important steps in curriculum construction. It is not necessary

here to point out that the selection of the educational objectives of a school and their validation require studies of several sorts. Valid educational objectives are not arrived at as a compromise among the various whims or preferences of individual faculty members but are reached on the basis of considered judgment utilizing evidence regarding the demands of society, the characteristics of students, the potential contributions which various fields of learning may make, the social and educational philosophy of the school or college, and what we know from the psychology of learning as to the attainability of various types of objectives. Hence, many of the schools spent a great deal of time on this step and arranged to re-examine their objectives periodically.

2. Classification of Objectives

As a second step, these statements of objectives from the Thirty Schools were combined into one comprehensive list and classified into major types. Before classification, the objectives were of various levels of generality and specificity and too numerous for practicable treatment. Furthermore, it was anticipated that the classification would be useful in guiding further curriculum development, because if properly made it would suggest types of learning experiences likely to be useful in helping to attain the objectives. A classification is of particular importance for evaluation because the types of objectives indicate the kinds of evaluation techniques essential to an adequate appraisal. The problem of classification is illustrated by the following partial list of objectives formulated by one school:

1. Acquiring information about various important aspects of nutrition
2. Becoming familiar with dependable sources of information relating to nutrition
3. Developing the ability to deal effectively with nutrition problems arising in later life

4. Acquiring information about major natural resources
5. Becoming familiar with sources of information regarding natural resources
6. Acquiring the ability to utilize and to interpret maps
7. Developing attitudes favoring conservation and better utilization of natural resources
8. Becoming familiar with a range of types of literature
9. Acquiring facility in interpreting literary materials
10. Developing broad and mature reading interests
11. Developing appreciation of literature
12. Acquiring information about important aspects of our scientific world
13. Developing understanding of some of the basic scientific concepts which help to interpret the world of science
14. Improving ability to draw reasonable generalizations from scientific data
15. Improving ability to apply principles of science to problems arising in daily life
16. Developing better personal-social adjustment
17. Constructing a consistent philosophy of life

These sample statements of objectives are of different levels of specificity and might well be grouped together under a smaller number of major headings. Thus, for purposes of evaluation, the several objectives having to do with the acquisition of information in various fields could be classified under one heading, since the methods of appraising the acquisition of information are somewhat similar in the various fields. Similarly, various objectives having to do with techniques of thinking, such as drawing reasonable inferences from data and the application of principles to new problems, could be classified under the general heading of development of effective methods of thinking, because the means of appraisal for these objectives are somewhat

similar. Furthermore, the methods of instruction appropriate for these techniques of thinking have similarities even though the content differs widely. Eventually, the following classification was used in general by the Staff:

MAJOR TYPES OF OBJECTIVES

1. The development of effective methods of thinking
2. The cultivation of useful work habits and study skills
3. The inculcation of social attitudes
4. The acquisition of a wide range of significant interests
5. The development of increased appreciation of music, art, literature, and other esthetic experiences
6. The development of social sensitivity
7. The development of better personal-social adjustment
8. The acquisition of important information
9. The development of physical health
10. The development of a consistent philosophy of life

This classification is not ideal but it served a useful purpose by focusing attention upon ten areas in which evaluation instruments were needed.³ It also helped to suggest emphases important in the curricular development of the Eight-Year Study. The classification of objectives will be improved as evidence accumulates regarding the social significance of different behavior patterns and regarding the correlation and consistency among the various specific reactions classified under each type of behavior. Until such research has been carried farther, each school or college will find useful some classification which serves the two purposes suggested.

³ The appraisal of the development of physical health, requiring, as it does, technical medical training, was not worked upon by the Evaluation Staff.

3. *Defining Objectives in Terms of Behavior* ^a

The third step was to define each of these types of objectives in terms of behavior. This step is always necessary because in any list some objectives are stated in terms so vague and nebulous that the kind of behavior they imply is not clear. Thus, a type of objective such as the development of effective methods of thinking may mean different things to different people. Only as "effective methods of thinking" is defined in terms of the range of reactions expected of students can we be sure what is to be evaluated under this classification. In similar fashion, such a classification as "useful work habits and study skills" needs to be defined by listing the work habits the student is expected to develop and the study skills which he may be expected to acquire.

In defining each of these classes of objectives, committees were formed composed of representatives from the Schools and from the Evaluation Staff. Usually, a committee was formed for each major type of objective. Since each committee included teachers from schools that had emphasized this type of objective, it was possible to clarify the meaning of the objective not in terms of a dictionary definition but rather in terms of descriptions of behavior teachers had in mind when this objective was emphasized. The committee procedure in defining an objective was to shuttle back and forth between general and specific objectives, the general helping to give wider implication to the specific, and the specific helping to clarify the general.

The resulting definitions will be found in subsequent chapters; however, a brief illustration may be appropriate here.^b The committee on the evaluation of effective methods of thinking identified various kinds of behavior which the Schools were seeking to develop as aspects of effective thinking. Three types of behavior patterns were considered important by all the Schools. These were: (1) the ability to formulate reasonable generalizations from specific data;

(2) the ability to apply principles to new situations; and (3) the ability to evaluate material purporting to be argument, that is, to judge the logic of the argument. When the committee proceeded to define the kinds of data which they expected students to use in drawing generalizations, the principles which they expected students to be able to apply, and the kinds of situations in which they expected students to apply such principles, and when they had identified the types of arguments which they expected students to appraise critically, a clear enough definition was available to serve as a guide in the further development of an evaluation program for this class of objectives. This process of definition had to be carried through in connection with each of the types of objectives for which an appraisal program was developed.

4. Suggesting Situations in Which the Achievement of Objectives Will Be Shown

The next problem was for each committee to identify situations in which students could be expected to display these types of behavior so that we could know where to go to obtain evidence regarding this objective. When each objective has been clearly defined, this fourth step is not difficult. For example, one aspect of thinking defined in the third step was the ability to draw reasonable generalizations from specific data. An opportunity to exhibit such behavior would be provided when typical sets of data were presented to students and they were asked to formulate the generalizations which seemed reasonable to them.

Another aspect of thinking defined in the third step was the ability to apply specified principles, such as principles of nutrition, to specified types of problems, such as those relating to diet. Hence, it seemed obvious that at least two kinds of situations would give evidence of such abilities. One would be a situation in which the student was presented

with these problems, for example, dietary problems, and asked to work out solutions utilizing appropriate principles of nutrition. Another kind of situation would be one in which the students were given descriptions of certain nutritional conditions together with a statement regarding the diet of the people involved, and the students were asked to explain how these nutritional conditions could have come about, using appropriate nutritional principles in their explanations.

As a third illustration, the definition of objectives identified as one educational goal the ability to locate dependable information relating to specified types of problems. It seemed obvious that a situation which would give students a chance to show this ability would be one in which they were asked to find information relating to these specified problems.

One value of this fourth step was to suggest a much wider range of situations which might be used in evaluation than have commonly been utilized. By the time the fourth step was completed, there were listed a considerable number of types of situations which gave students a chance to indicate the sort of behavior patterns they had developed. These were potential "test situations."

5. *Selecting and Trying Promising Evaluation Methods*

The fifth step in the evaluation procedure involved the selection and trial of promising methods for obtaining evidence regarding each type of objective. Before attempting to construct new evaluation instruments, each committee examined tests and other instruments already developed to see whether they would serve as satisfactory means for appraising the objective. Only limited test bibliographies were then available.⁴ In addition to examining bibliographies, the

⁴ Now, any group working on an evaluation program will find useful a more complete bibliography of evaluation instruments, such as the *Buros Mental Measurements Yearbook*. This bibliography not only lists tests and other appraisal instruments which are commercially available, but also in-

committees obtained copies of those instruments which seemed to have some relation to their objectives. In examining an instrument the committee members tried to judge whether the student taking the test could be expected to carry out the kind of behavior indicated in the committee's definition of this objective. Then, too, the situations used in the instruments were compared with those suggested in the fourth step as to their likelihood of evoking the behavior to be measured. The committees recognized that they might be misled by undue optimism in the name or the description of the test, and sought to guard against it. Even though a test was called a general culture test, or a world history test, or a general mathematics test, it was generally found that it measured only one or two of the objectives which teachers of these fields considered important. In order to estimate what the test did measure, it was necessary to examine the test situations to judge what kind of reaction must be made by the student in seeking to answer the questions. It also proved useful to examine any evidence reported which helped to indicate the kind of behavior the test was actually measuring.

At this point most of the committees found that no tests were available to measure certain major aspects of the important objectives. In such cases, it was necessary to construct additional new instruments in order to make a really comprehensive appraisal of the educational program in the Thirty Schools. The nature of the instruments to be built varied with the types of objectives for which no available instruments were found. Every committee, however, found it helpful in constructing these instruments to set up some of the situations suggested in step four and actually to try them out with students to see how far they could be used as

cludes several critical reviews of these tests written by teachers, curriculum constructors, and test makers. These reviews help in selecting from available instruments those which might be worth a trial.

test situations. By the time the fifth step had been carried through, certain available tests were selected and tried out and certain new appraisal instruments were constructed and given tentative trial.

6. *Developing and Improving Appraisal Methods*

The sixth major step was to select on the basis of this preliminary trial the more promising appraisal methods for further development and improvement. This further development and improvement was largely the responsibility of the Evaluation Staff. The committees met from time to time to review the work of the Staff, and many teachers were asked to criticize and make suggestions for improvement. Obviously, however, the detailed work had to be done by the Staff.

The basis for selecting devices for further development included the degree to which the appraisal method was found to give results consistent with other evidences regarding the student's attainment of this objective and the extent to which the appraisal method could be practicably used under the conditions prevailing in the Schools. The refinement and improvement consisted in working out directions which were unambiguous, modifying exercises which were found not to give discriminating results, eliminating exercises which were found to be almost exact duplicates of other exercises in terms of the type of reaction elicited from the student, developing practicable and easily interpretable records of the student's behavior, and making other revisions which gave more clear-cut measures, which provided a more representative and adequate sample of the student's reaction, and which improved the ease with which the instrument could be used.

An important problem in the refinement and improvement of an evaluation instrument proved to be the determination of the aspects of student behavior to be summarized

and the decision regarding the units or terms in which each aspect was to be summarized. For example, consider a test constructed to appraise the ability of students to formulate reasonable generalizations from data new to them. An obvious type of test situation would be one in which sets of data new to the student were presented to him and he was asked to examine the data and to formulate generalizations which seemed reasonable to him. When we approach the question of summarizing his behavior in some form which provides a measurement or appraisal, we are faced with the problem of identifying aspects, that is, dimensions of the behavior to measure, and of deciding upon units of measurement to use. One aspect which is important in judging the value of the generalization formulated is its relevance. Generalizations which have no relevance to the data are obviously not satisfactory. If this aspect is to be measured, there are several possible units of measurement which might be used. For example, we could set up a subjective scale for degree of relevance and have judges apply this scale to each generalization, rating it at some point on this scale. Another unit of measurement could be used by classifying each generalization as relevant to the data or irrelevant to the data, thus measuring the relevance in terms of the number of the student's generalizations which are classified as relevant. On the other hand, since students may differ markedly in the total number of generalizations formulated, a better unit of measure for the degree of relevance might be the per cent of the student's generalizations which are classified as relevant.

A second aspect which has some importance in appraising generalizations of this type would be the degree to which relevant generalizations are carefully formulated and involve no overgeneralizations, that is, generalizations more sweeping than the data would justify. If this aspect were chosen as part of the appraisal, several possible units could be used in the measurement. One possible unit might be the

judgment of the reader of the paper as to the degree to which each generalization was carefully or incautiously formulated. This kind of unit involves a considerable degree of subjective judgment so that many might prefer the simple categorization of each relevant generalization as either going beyond the data or not going beyond the data. In this case, a unit of measurement might be the per cent of relevant generalizations not going beyond the data. Perhaps these illustrations are sufficient to show that it is always necessary in the development of new evaluation instruments or in the use of those which have been developed by others to decide on the aspects of the behavior to be described or measured and the terms or units which will be used in describing or measuring this behavior.

7. Interpreting Results

The seventh and final step in the procedure of evaluation was to devise means for interpreting and using the results of the various instruments of evaluation. The previous steps resulted in the selection or the development of a range of procedures which could be used periodically in appraising the degree to which students were acquiring the objectives considered important in a given school. These instruments provided a series of scores and descriptions which served to measure various aspects of the behavior patterns of the students. As these instruments were used, a great number of scores or verbal summaries became available at each appraisal period. Each of these scores or verbal summaries measured an aspect of behavior considered important and represented a phase of the objectives of the school. The Staff then conducted comparability studies for certain of the instruments so that the scores or verbal summaries could be compared with scores or verbal summaries previously obtained; by this comparison some estimate of the degree of change or growth of students could be made. However, the

meaning of these scores became fuller through various additional studies.

One type of study involved the identification of scores typically made by students in similar classes, in similar institutions, or with other similar characteristics. Another helpful study involved a summary and analysis of the typical growth or changes made in these scores from year to year. A third type involved studies of the interrelationship of several scores to identify patterns. These patterns are not only useful when obtained among several scores dealing with the behavior relating to one objective, but are also useful in seeing more clearly the relation among the objectives. It was pointed out in the introductory section of this chapter that human behavior is to a large degree unified and that efforts to analyze behavior into different types of objectives are useful but may do some harm if the essential interrelationships of various aspects of behavior are forgotten. It was found important in this seventh step to examine the progress students were making toward each of the several objectives in order to get more clearly the pattern of development of each student and of the group as a whole and also to obtain hypotheses helpful in explaining the types of development taking place. Thus, for example, the evaluation results in one school showed that students were making marked progress in the acquisition of specific information and were also shifting markedly in their attitudes toward specific social issues, but at the same time they showed a high degree of inconsistency among their various social attitudes, and were making little progress in applying the facts and principles learned. These results suggested the hypothesis for further study that the students were being exposed to too large an amount of new material and were not being given adequate opportunity to apply these materials, to interpret them thoroughly, and to build them into their previous ideas and beliefs. A test of this hypothesis was made by modifying the

course so as to provide for a smaller amount of new material, the introduction of more opportunities for application, and the emphasis upon thoroughness of interpretation and reorganization. This revision in the course resulted in corresponding improvements in the pattern of student achievement. If this revision had not resulted in corresponding improvements, other hypotheses which might explain the results would have been considered. This procedure illustrates a useful means of interpreting the results of several evaluation instruments. It was found that each school needed methods for interpreting and using the results of appraisal so as to improve the educational program and to guide individual students more wisely.

The usefulness of the evaluation program depends very largely upon the degree to which the results are intelligently interpreted and applied by the teachers and school officers. The Evaluation Staff, however, had some responsibility in developing methods for interpreting the results intelligently and in helping teachers and school officers to use them most helpfully. Hence, in addition to making these studies of the instruments, members of the Evaluation Staff visited a number of the Schools and went over the results with the school staffs, suggesting possible interpretations and indicating methods by which these interpretations could be more adequately verified and used. As a result of these preliminary visits, certain methods of interpretation were developed. At this point members of the school staffs who were participating in summer workshops were asked to try these methods of interpretation and to criticize them. Then, for a period of two years, opportunity was provided for at least one representative from each school to spend a considerable period of time in the staff headquarters to gain further familiarity with the evaluation instruments, with their interpretation, and with their use. These school representatives received the training on the assumption that they would have oppor-

tunity for giving leadership to the evaluation program in their respective schools. As a result of this experience, the staff believes that a program of testing or evaluation can reach greater fruition when a systematic attempt is made to provide for the training of teachers and school officers in the interpretation and use of evaluation results.

DIVISION OF LABOR IN THE EVALUATION PROGRAM

The previous description of the development of the evaluation program explained that it involved the cooperation of the school personnel and the Evaluation Staff. This does not imply that teachers, school officers, and Evaluation Staff members were all performing the same functions. Although there was some overlapping of functions, there was also a general plan for division of labor. One major division of labor was based on the principle that the school's duty is to evaluate its program, while the technician's function is to help develop means of evaluation. Furthermore, in following through the steps of evaluation, there was some division of duties. Every faculty member and school officer bore some responsibility for the formulation of the objectives of his school. The classification of objectives into major types of behavior was largely a function of the Evaluation Staff because the primary purpose of this classification was to place in the same group those objectives which involved similar types of student reactions, and which might conceivably involve somewhat similar techniques of appraisal.

The further definition and clarification of each class of objectives was the task of an interschool committee composed of teachers, school officers, and members of the Evaluation Staff. The staff members raised questions and suggested directions for discussion which would help to define or clarify the given type of objective, but most of the defining was done by the representatives of the schools which had emphasized this type of objective.

The interschool committee also suggested situations in which the desired behavior might be shown by students. The school representatives then assumed responsibility for trying out these situations to see if they would serve as means of evaluation. The review of these trials, their criticism, and plans for improving the methods of evaluation were carried on by the entire committee. From this point on, the refining of the evaluation instrument and its development for constructive use was largely the task of members of the Evaluation Staff. However, teachers and school officers gave helpful criticisms and suggestions and eventually determined whether an instrument was worth using and could practically be used in a given school. Finally, the school staff was expected to assume responsibility for obtaining evidence of growth and studying these results.

This plan has wide applicability. It provides a way in which technicians in testing and evaluation may work constructively with teachers and school officers to develop an evaluation program. It avoids the danger on the one hand of having instruments constructed by technicians who are not clear about the curriculum and guidance program of the school, and on the other hand the formulation of an evaluation program by persons who are relatively unfamiliar with methods of describing and measuring human behavior.

SUMMARY

This brief description of the steps followed in developing the evaluation program should have indicated that the process of evaluation was conceived as an integral part of the educational process. It was not thought of as simply the giving of a few ready-made tests and the tabulations of resulting scores. It was believed to be a recurring process involving the formulation of objectives, their clearer definition, plans to study students' reactions in the light of these objectives, and continued efforts to interpret the results of

such appraisals in terms which throw helpful light on the educational program and on the individual student. This sort of procedure goes on as a continuing cycle. Studying the results of evaluation often leads to a reformulation and improvement in the conception of the objectives to be obtained. The results of evaluation and any reformulation of objectives will suggest desirable modifications in teaching and in the educational program itself. Modifications in the objectives and in the educational program will result in corresponding modifications in the program of evaluation. So the cycle goes on.

As the evaluation committees carried on their work, it became clear that an evaluation program is also a potent method of continued teacher education. The recurring demand for the formulation and clarification of objectives, the continuing study of the reactions of students in terms of these objectives, and the persistent attempt to relate the results obtained from various sorts of measurement are all means for focusing the interests and efforts of teachers upon the most vital parts of the educational process. The results in several schools indicate that evaluation provides a means for the continued improvement of the educational program, for an ever deepening understanding of students with a consequent increase in the effectiveness of the school.

The subsequent chapters describe in more detail the development of evaluation instruments for certain types of objectives. Space does not permit the description of all the evaluation instruments developed. Tests of effective methods of thinking are described because this objective was of concern to all the schools, and few instruments of this sort had previously been developed. On the other hand, although work habits and study skills were emphasized in most of the schools, the description of the instruments developed is not included in this report. The committee identified the fol-

lowing work habits and study skills for which methods of appraisal were needed:

Range of Work Habits and Study Skills

- 1.1 Effective Use of Study Time
 - 1.11 Habit of using large blocks of free time effectively
 - 1.12 Habit of budgeting his time
 - 1.13 Habit of sustained application rather than working sporadically
 - 1.14 Habit of meeting promptly study obligations
 - 1.15 Habit of carrying work through to completion
- 1.2 Conditions for Effective Study
 - 1.21 Knowledge of proper working conditions
 - 1.22 Habit of providing proper working conditions for himself
 - 1.23 Habit of working independently, that is, working under his own direction and initiative
- 1.3 Effective Planning of Study
 - 1.31 Habit of planning in advance
 - 1.32 Habit of choosing problems for investigation which have significance for him
 - 1.33 Ability to define a problem
 - 1.34 Habit of analyzing a problem so as to sense its implications
 - 1.35 Ability to determine data needed in an investigation
- 1.4 Selection of Sources
 - 1.41 Awareness of kinds of information which may be obtained from various sources
 - 1.42 Awareness of the limitations of the various sources of data
 - 1.43 Habit of using appropriate sources of information, including printed materials, lectures, interviews, observations, and so on
- 1.5 Effective Use of Various Sources of Data
 - 1.51 Use of library
 - 1.511 Knowledge of important library tools

- 1.512 Ability to use the card catalogue in a library
- 1.52 Use of books
 - 1.521 Ability to use the dictionary
 - 1.522 Habit of using the helps (such as the Index) in books
 - 1.523 Ability to use maps, charts and diagrams
- 1.53 Reading
 - 1.531 Ability to read a variety of materials for a variety of purposes using a variety of reading techniques
 - 1.532 Power to read with discrimination
 - 1.533 Ability to read rapidly
 - 1.534 Development of a more effective reading vocabulary
- 1.54 Ability to get helpful information from other persons
 - 1.541 Ability to understand material presented orally
 - 1.542 Facility in the techniques of discussion, particularly discussions which clarify the issues in controversial questions
 - 1.543 Ability to obtain information from interviews with people
- 1.55 Ability to obtain helpful information from field trips and other excursions
- 1.56 Ability to obtain information from laboratory experiments
- 1.57 Habit of obtaining needed information from observations
- 1.6 Determining Relevancy of Data
 - 1.61 Ability to determine whether the data found are relevant to the particular problem
- 1.7 Recording and Organizing Data
 - 1.71 Habit of taking useful notes for various purposes from observations, lectures, interviews, and reading
 - 1.72 Ability to outline material for various purposes
 - 1.73 Ability to make an effective organization so that the material may be readily recalled, as in notetaking

- 1.74 Ability to make an effective organization for written presentation of a topic
- 1.75 Ability to make an effective organization for oral presentation of a topic
- 1.76 Ability to write effective summaries
- 1.8 Presentation of the Results of Study
 - 1.81 Ability to make an effective written presentation of the results of study
 - 1.811 Habit of differentiating quoted material from summarized material in writing reports
 - 1.812 Facility in handwriting or in typing
 - 1.82 Ability to make an effective oral presentation of the results of study
- 1.9 Habit of Evaluating Each Step in an Investigation
 - 1.91 Habit of considering the dependability of the data obtained from various sources
 - 1.92 Habit of considering the relative importance of the various ideas obtained from various sources
 - 1.93 Habit of refraining from generalization until data are adequate
 - 1.94 Habit of testing his own generalizations
 - 1.95 Habit of criticizing his own investigations

A number of preliminary instruments were constructed for this extensive list of habits and skills.⁵ Most of these have not been sufficiently refined to justify inclusion in this volume.

Instruments for appraising social attitudes are treated in the chapter on the evaluation of social sensitivity. Because so many tests of information were already available, and because techniques for measuring the recall and use of information were well understood by teachers, the committees did not devote major attention to developing further instru-

⁵ A monograph, "Study Skills and Work Habits: Some Selected Materials," was prepared by a committee headed by Cecile White Flemming of the Horace Mann School for Girls, and was circulated in mimeographed form in 1935. It is now out of print.

ments of this type. A few were constructed for specific purposes, but these are not reported here.

The appraisal of the philosophy of life developed by the students involves the use of evidence from many of the other areas, such as thinking, social attitudes, interests, appreciations, and social sensitivity. Hence, methods for evaluating the student's philosophy of life are primarily methods of combining and interpreting the results of other measurements. Methods of interpretation are discussed in Chapter VII. Finally, the planning of a comprehensive evaluation program and the problems of recording are considered.

It is obvious that there are other areas and other problems in the construction and use of evaluation instruments still untouched. The Evaluation Staff hopes, however, that its experience will be useful in guiding further endeavor so that ultimately schools may be able to evaluate their work with a high degree of comprehensiveness.

Chapter II

ASPECTS OF THINKING



INTRODUCTION

The responsibility of secondary schools for training citizens who can think clearly has been so long and so frequently acknowledged that it is now almost taken for granted. The educational objectives classifiable under the generic heading "clear thinking" are numerous and varied as to statement, but there can be little doubt concerning their fundamental importance. Although in recent years there has been increasing recognition of other responsibilities and purposes, there has been little accompanying tendency to demote clear thinking to a minor role as an educational objective. It was therefore not surprising to find considerable emphasis upon this objective in the statements of purposes submitted to the Evaluation Staff by the schools participating in the Eight-Year Study.

The fact that an objective has been stated frequently or with emphasis does not insure that its meaning and implications are sufficiently clear to guide effective teaching or to serve as a basis for the evaluation of achievement. In this respect the "clear thinking" objectives as originally stated by the schools were no different from other even more "intangible" objectives. An examination of the pertinent educational literature, moreover, revealed that most of the available analyses of these objectives were unsatisfactory for the purpose of evaluation. It therefore proved necessary to devote considerable time to clarification of the objectives and to analysis of the behaviors which would reveal that students

were achieving them. In the course of the analysis it was convenient to break up the general objective into a limited number of component parts, and then to analyze each of these in some detail. The aspects of clear or "critical" thinking which were selected dealt with the ability to *interpret data*, with the ability to *apply principles* of science, of the social studies, and of logical reasoning in general, and finally, with certain abilities associated with an understanding of the *nature of proof*. This chapter will be devoted chiefly to the description of each of these aspects as they were eventually analyzed, and to a description of some of the evaluation instruments which were developed to evaluate the associated abilities.

It may be well to note at the outset that the abilities involved in the aspects of thinking listed above are overlapping. Although the abilities called into action in a successful interpretation of a set of data seem to be primarily inductive, and those utilized in the other aspects are more deductive in nature, it is neither necessary nor desirable to emphasize such distinctions. In connection with any given problem, the process of reflective thinking, as defined by Dewey and others, is likely to call upon a number of the abilities to be described in connection with each major aspect of thinking mentioned above. It should also be noted that other important aspects of thinking—for example, the ability to formulate hypotheses—are only implicitly included in the above list and receive only cursory attention in the following discussion. The separation of clear thinking into these and other aspects is a product of the analysis and is not to be considered as inherent in the process of clear thinking. It was convenient because it facilitated the exploration of the larger objective and the development of practicable means of evaluation. A satisfactory evaluation of the thinking abilities of students involves a synthesis of the data obtained from various instruments.

The four major aspects of clear thinking³ listed above not only overlap among themselves, but they also overlap with other educational objectives. The attitudes and the emotions of students may influence their ability to think clearly in certain situations. This has been explicitly recognized in the analyses of these objectives and in the construction of the evaluation instruments to be described in this chapter. At the moment, it is necessary to mention only that evaluation of the *disposition* to think critically has not been extensively worked upon and is not discussed in the following pages. In the opinion of the Evaluation Staff, the best available means is some sort of observational record, and this method demands only the simplest of techniques supported by alert sensitivity and perseverance on the part of the observer. Evidence of the disposition to think critically collected by this method would, however, be a valuable addition to other evidence relevant to clear thinking of the sort to be described later.

The scope of this phase of the evaluation project made it necessary to omit many details in the discussion of some of the instruments. For purposes of illustration, certain procedures are explained at length in relation to a selected instrument, and are condensed or omitted elsewhere. The analysis of the application of principles in the field of social science is treated somewhat differently from that for the natural sciences, and will consequently be found in Chapter III on "Social Sensitivity." The following sections include the analyses which were made of the ability to interpret data, of the application of principles of science and of logical reasoning, and of abilities associated with an understanding of the nature of proof. The instruments to measure achievement that were developed and some of their technical characteristics and uses will also be described. No account is given of similar instruments developed by individual teachers.

I. INTERPRETATION OF DATA

ANALYSIS OF THE OBJECTIVE

The Committee on the Interpretation of Data, composed of representatives from each school interested in this objective and members of the Evaluation Staff, began with two major questions: What do students do when they interpret data well? What kinds of data should they be able to interpret?

Behaviors Involved in Interpretation of Data

Some conceived of interpretation as a complex behavior which included the ability to judge the accuracy and relevance of data, to perceive relationships in data, to recognize the limitations of data, and to formulate hypotheses on the basis of data. From the wide range of behaviors which were suggested, the committee selected two which seemed to them to be of paramount importance: (1) the ability to perceive relationships in data, and (2) the ability to recognize the limitations of data.

The first of these involves the ability to make comparisons, to see elements common to several items of the data, and to recognize prevailing tendencies or trends in the data. These behaviors are dependent on the ability to read the given data, to make simple computations, and to understand the symbolism used. It became apparent that these operations vary for different types of data. Thus in the case of graphic presentation the student must be able to locate specific points on the graph, relate these to the base lines, recognize variations in length of bars or slope of graph line, and so on. In many cases, students must understand simple statistical terms (e.g., "average"), the units used, and the conventional methods of presentation of different forms of data.

A second type of behavior which the teachers expect of students is the ability to recognize the limitations of given data even when the items are assumed to be dependable. A

student who develops this ability recognizes what other information, in addition to that given, he must have in order to be reasonably sure of certain types of interpretations. He refrains from making judgments relative to implied causes, effects, or purposes until he has necessary facts at hand. He recognizes the error in allowing his emotions to carry him beyond the given facts when he judges conclusions that affect him personally. If he holds rigidly to what is established by the data, the kinds of generalizations that he can make without qualifications are limited. He recognizes that many interpretations must be regarded as almost completely uncertain because the facts given are insufficient to support such interpretations even with appropriately stated qualifications.

These behaviors do not preclude the possibility of making qualified inferences when the situation warrants. This type of interpretation can be made, for example, when the data reveal definite trends. By qualifying the statement with words such as "probably" a student may then *extrapolate*, that is, make interpretations which are somewhat beyond the facts but in agreement with a definitely established trend. Or a student may *interpolate*, in other words, make a qualified inference concerning an omitted point between observed points in a set of data which reveal an established trend. In another case, a student may risk a qualified prediction relative to similar sets of data applying to similar conditions. Even when the inferences are qualified, the student must be careful not to allow his statements to go far beyond the observed facts. These inferences are necessarily confined to a rather narrow range whose extent depends somewhat on the subject to which the data apply. Fundamentally, the objective involves making a distinction between what is established by the data alone, and what is being read into the data by the interpreter.

During the analysis of the objective it was also recognized

that the ability to make original interpretations and the ability to judge critically interpretations made by others might not be closely related. When judging a stated interpretation one may derive a clue that directs attention to specific relationships in the data. An original interpretation usually involves the ability to perceive these relationships without the aid of suggestions or directions. In the discussion of this point it was noted, on the one hand, that relatively few individuals have occasion to collect data and make original interpretations, since most of the data encountered in life are already wholly or partially interpreted. Critical judgment of these interpretations is, however, very important. On the other hand, it was noted that some individuals do have frequent need to collect data and formulate original interpretations, and almost everyone has some need of the abilities involved. A decision was made to concentrate primarily upon evaluation of the ability to judge interpretations made by others, and to study the relationship between this and the ability to make original interpretations.

Several other behaviors were recognized as ones which may be considered important in connection with the interpretation of data. One of these is the ability to evaluate the dependability of data; another is the ability to formulate hypotheses. In evaluating the dependability of data, a student might question the competence, bias, or integrity of the person who presents the data; he might attempt to determine the adequacy and appropriateness of the methods, techniques, and controls used in obtaining the data; he might question the adequacy and the appropriateness of the methods of summarizing the data. In formulating hypotheses on the basis of given data, the student might infer probable causes or he might predict probable effects. Information other than that given in the data may be required in order to make a satisfactory evaluation or to formulate a reasonable

hypothesis. Thus recall of information might also be regarded as an ability involved in the interpretation of data.

Although the importance of all these aspects of interpretation of data was fully recognized, the teachers selected for more intensive study those behaviors on which they proposed to give the greatest emphasis in their respective schools. Whether a student is making original interpretations or judging interpretations made by others, the teachers expect the student who has achieved the objective to perceive relationships in data and to recognize the limitations of data. These two important behaviors were therefore selected for particular attention in developing evaluation instruments.

Kind of Data

The second major question which had to be answered in analyzing the objective dealt with the kinds of data that students should be able to interpret. The committee recognized several different ways of classifying data. Among these were the following: (1) according to the form of presentation, (2) according to the subject-matter fields from which the data are drawn, (3) according to problems or areas of living with which the data deal, (4) according to types of relationships inherent in the data, (5) according to the purpose the data are intended to serve, (6) according to various levels of generality, (7) according to the degree to which the possibility of making meaningful interpretations depends upon the knowledge of other facts.

The form of presentation of data may vary. For example, data may be presented in graphical form. Pictures, maps, cartoons, and various types of graphs, such as line or bar graphs, are familiar examples. Data also are often presented in tabular form. Such tables are frequently found in reports of experiments, election returns, scores of baseball games, and so on. Sometimes data are not set off from the prose form of reading matter but are incorporated in the context.

This method of presentation is often used in editorials, printed speeches, and news items. Sometimes the same data are presented in several forms; this situation is commonly found in advertisements, for example.

Data may be drawn from various subject fields. Data from the fields of economics and sociology commonly appear in newspapers, magazines, and current books. Data from the fields of physics, chemistry, biology, and other sciences are presented in many publications which are commonly read; advertisements, for example, often incorporate data from these fields.

The classification of data in terms of areas of living or problems would probably make use of categories such as vocation, health, government, transportation, family relationships, and others of similar type. Classification according to types of relationship would emphasize categories such as chronological trends, relationship of parts to a whole, and the like. If data are differentiated in terms of the purposes which they are intended to serve, distinctions may be made, for example, between what purports to be an impartial presentation of facts and a presentation intended to sell a particular idea or defend a special interest. Different levels of generality are illustrated by data showing unemployment in a single city in contrast to data on unemployment in an entire state or country. If the latter are available, often more meaningful interpretations could be made concerning the situation in the single city, and hence this same illustration indicates how additional information may influence the interpretation, and how the amount of such information needed may form a basis of classification.

Although other classifications are possible and were considered, for purposes of evaluation the teachers chose the following criteria for the selection of the data to be presented to students for interpretation: (1) data presented in various forms; (2) data relating to various subject fields; (3) data

relating to major problem areas; (4) data including various types of relationships. As is often the case, these criteria are not independent, and a given set of data will satisfy several criteria simultaneously.

In order that the interpretation may not be made from memory, it is necessary that the data be "new" to the student in the sense that this particular organization of the facts has not previously been interpreted for the student by someone else. If he has heard or read an interpretation of this organization of facts, his response may represent recall of an interpretation made by another and not give a measure of his own ability to interpret.

The analysis of the objective thus resulted not only in a description of the behaviors which might be included under the phrase "interpretation of data," but also in a conscious restriction of the scope of the eventual evaluation. This restriction applied to the types of behavior which were to be emphasized, and to the criteria for the selection of data which were to be presented to students.

THE DEVELOPMENT OF EVALUATION INSTRUMENTS

Preliminary Investigations

Observations of a student's many overt behaviors in responding to data of various kinds is one way in which evidence of his ability to interpret data may be obtained. This type of evidence can probably be best secured by observational records kept by teachers or other persons trained to observe and record these behaviors. Under certain conditions a student's written materials, such as laboratory notebooks, papers, etc., may be a fruitful source of evidence. However, the time consumed and the possible lack of objectivity of scores present serious difficulties in the use of these techniques. Since these methods usually involved more or less uncontrolled situations, teachers were interested in devising a method that would better stabilize some of the

variable factors. The method which was selected makes use of pencil and paper tests in which the student reacts in writing to written data. Many methods of obtaining this type of evidence have been experimented with in the Study. A few will be discussed to present some of the approaches used and some of the difficulties the Evaluation Staff has encountered in measuring the abilities involved. One of the most direct methods used was to present the student with sets of written data, ask him to write true statements concerning the data, and to appraise the interpretations which he wrote. However, such a free-response essay-form presents several difficulties in evaluation. It was found that even when the number of interpretations to be made is specified in the directions, individual students tend to use a narrow range of relationships in their responses. Thus, the responses do not adequately sample the types of interpretations which the students are capable of making when their attention is focussed on data relating to their own particular problems or concerns, or when breadth of treatment is encouraged by more specific directions in the test. Moreover, great difficulty is experienced in scoring such a test, for it is often impossible to be reasonably sure what the student means by his written statements. This perplexity may arise from ambiguity or incompleteness of student's statements or from peculiarities in his style. It is possible to attain high objectivity for such a test, but only after elaborate criteria for scoring have been carefully set up. Even with such a device, it is a time-consuming method. In one case, for example, it required approximately 90 hours for each of the trained markers to score 193 papers of ten exercises calling for responses of this type. Because of these difficulties, this method of getting evidence of a student's ability to interpret data is impractical for most teachers.

In order to determine the types of interpretations students should be expected to judge critically and the kinds of errors

commonly made in interpreting data, a study was made of interpretations commonly found in editorials, advertisements, news items, reports of scientific experiments, and similar materials. For instance, the conclusions of many reports of experiments were critically studied in relation to the data on which they were based. In this and other such studies it was possible to discover the kinds of relationships involved in the interpretations, the kind of assumptions that were made, the accuracy and adequacy of the inferences made from the data. When students' essay responses were also critically studied in the same way and comparisons made, it became apparent that the interpretations from both these sources were susceptible to virtually the same types of classifications. One classification that could be made was in terms of the kind of relationships involved. For convenience of reference, these types are denoted by various words or phrases, such as "extrapolation," "comparison of points," or "cause." They are as follows:¹

1. *Reading Points.* This type of statement is usually merely a restatement of the data.
2. *Comparison of Points.* The statement is a comparison of two or more items or "points" in the data.
3. *Cause.* The statement presents a cause of conditions presented in the data.
4. *Effect.* The statement formulates a prediction of a probable effect of the conditions described.
5. *Value Judgment.* The statement presents a recommended course of action suggested by the data, or an opinion of what ought to be or ought not to be.
6. *Recognition of Trend.* The statement describes a prevailing tendency or trend in the data.
7. *Comparison of Trends.* The statement presents a

¹ For examples of statements of these types, see the sample problem on page 52.

comparison of two or more prevailing tendencies or trends in the data.

8. *Extrapolation*. The statement formulates a prediction of a point or item or fact which is not given in the data and lies beyond points or items or facts which are given in the data.
9. *Interpolation*. The statement formulates a prediction of a point or item or fact of data which lies between points or facts which are given in the data.
10. *Sampling*. The statements concern (a) only a part of the group described in the data, or (b) a larger group containing as a part of itself the group described in the data.
11. *Purpose*. The statement presents a judgment of purpose of the given data.

These types of interpretations may be also arranged into a concise and meaningful classification which emphasizes the difference in degree of accuracy with which they are used by students. Thus, students' responses may include the following:

1. Interpretations which are *accurate*. These interpretations may formulate comparisons, trends, and specific facts which are established by the data as true or false and are correctly stated without qualification. Other interpretations under this classification may be concerned with sampling, extrapolation, or interpolation. They are not fully supported by the given data, but are probably true or probably false on the basis of the trends established in the data, and are stated by the student with sufficient qualification.
2. Interpretations which are *overgeneralizations*—that is, interpretations containing unqualified or unwarranted statements involving interpolation, extrapolation, and sampling, or statements of cause, purpose,

effect which cannot be established by the given data even in qualified form. This type of error may be referred to as "going beyond the data."

3. Interpretations which are undergeneralizations—that is, which involve unnecessarily qualified statements concerning specific facts, trends, and comparisons which are established in the data. Such departures from accuracy may be referred to as "overcaution."
4. Interpretations which involve "crude errors"; for example, the student errs by misreading the points or trends in the data, by failing to understand meanings of terms, such as "average" and "per cent," or by failing to relate properly the data of a graph to the base lines.

Such analyses provided a basis for construction of a short-answer type of test exercise. This type of test does not present the difficulties in scoring inherent in the essay form and makes it possible for a student to react to many types of data in a limited time. During its development, the short-answer test has passed through several transitional forms. Analysis and statistical study of early forms suggested changes which were incorporated in subsequent forms. For the sake of simplicity of explanation, only the latest form of the interpretation of data test (Form 2.52) will be described in detail.

Structure of Interpretation of Data Test, Form 2.52

The test to be described is intended primarily for the senior high school level. It contains ten sets of data selected to satisfy the criteria set up by the committee interested in the objective. These data are presented in various forms, including tables, prose charts, and different kinds of graphs. The problems are selected from several fields (such as medi-

² Overcaution is not considered an error by everyone. Some consider it evidence of a tendency to suspend judgment until further evidence is available.

cine, home economics, sociology, genetics) and contain data pertinent to such topics as technological unemployment, heredity, crop rotation, immigration, government expenditures, and health.

Each set of data is followed by 15 statements which purport to be interpretations. The student is asked to indicate his judgment of each of the statements by placing it in one of five categories as indicated by the short code given at the top of the sample exercise on page 52. In the sample, the list of responses accepted as correct by a jury of competent persons is given in the margin before each interpretation. A word or phrase describing the main kind of relationship involved follows each interpretation.

A study of the sample exercise in relation to the following summary of the procedure used in constructing the test will indicate how the analyses described previously were utilized. It may also serve as a guide for teachers who wish to construct similar tests suited for use with their own students.

1. The data were selected according to the criteria set up by the committee.
2. Fifteen interpretative statements were made from each set of data. The types of statements included were based on an analysis of types of interpretations which were found in current literature, the judgment of teachers who were concerned with the objective, and the analysis of responses of students who were asked to write original interpretations. This approach was used both to give the students an opportunity to judge statements including typical errors made in interpretations, and to insure the inclusion in the test of types of interpretations which students encounter and are capable of recognizing. These interpretations involve the following types of behaviors: comparisons of points of data, recognition and comparison of

- trends, judgments of cause, effect,³ purpose, value, analogy,⁴ extrapolation, interpolation, and sampling.
3. The types of relationship involved in the interpretations which the students are asked to judge were distributed among the five response categories as follows:

- a. Interpretations adequately supported by the data, and so worded that they are meant to be judged by the students as *true*. These statements require the student to judge interpretations that involve: comparison of points in the data; recognition of trends; and comparison of trends. Ten per cent of the total number of statements in the test are in this category.⁴
- b. Interpretations inadequately supported by the data, so worded that they are meant to be judged *probably true*. These statements require the students to judge interpretations that involve a knowledge of the principles of prudent extrapolation, interpolation, and sampling as previously defined. They include inferences that go beyond the data but are suggested by the data and are based on trends or facts in the data. They also include some conclusions that would be popularly interpreted as true. They are intended to contribute information concerning the ability of students to recognize the necessity for qualification in interpretation. About 20 per cent of the total number of statements are in this category.
- c. Interpretations inadequately supported by the

³ Although in this Study analogy was not found to be used to any great extent in student-written interpretations of data, this type of interpretation is encountered extensively in advertising, newspaper articles, etc. It was also the thought of the Evaluation Staff that analogy is one aspect of scientific thinking which they desired to measure in several different contexts. It appears also in the Application of Principles of Science tests.

⁴ This distribution was based upon studies of reliabilities of early forms.

data, so worded that they are meant to be judged as based upon *insufficient data*. They give opportunity for the student to make judgments concerning statements of analogies relating to the data, concerning statements referring to a cause or an effect of the situation revealed by the data, concerning the purpose the data are supposed to serve, and concerning a recommended course of action supposedly desirable on the basis of the data. Also included are some statements depending upon an injudicious use of interpolation, extrapolation, and sampling. About 40 per cent of the total number of statements are in this category.

- d. Interpretations inadequately supported by the data, so worded that they are meant to be judged *probably false*. These include inferences which are suggested by the data but which are contrary to the trends of facts in the data, and conclusions which would be popularly interpreted as false. The same types of interpretations are used here as in *b*. Twenty per cent of the total number of statements are in this category.
 - e. Interpretations which are contradicted by the data, so worded that they are meant to be judged as *false*. These statements involve the same types of interpretations as are listed in *a* above. Ten per cent of the total number of statements are in this category.
4. Within each test exercise the interpretations were arranged in random order. Directions to the students were formulated. These directions asked students to place each statement in one of the five different categories.

Before the test was considered ready for use, an analysis

of student responses was made. In each case where the judgment of a large number of students conflicted with the key, there was an attempt to analyze the student's thinking to see if the conflict in judgment was due to confusion in the test or to an erroneous concept held by the students. Ambiguous statements were revised, and a final key was drawn up. The scores made by students are, therefore, to be considered as a means of comparison of their thinking with the judgments of the jury.

Summarization of Scores

For purposes of exposition, the manner in which the answer sheets from a class are scored may be described as follows. By tabulating a student's response for each item in relation to the jury's key for that item in the proper cell of the following chart, a teacher can describe student's achievement both as to accuracy and as to errors.⁵

As indicated by the chart, student responses can be described in the following terms: *general accuracy*, *caution*, *beyond data*, and *crude errors*. This terminology may be defined as follows: *General accuracy* means the extent to which the student agrees with the jury—that is, recognizes true statements as *true*, probably true as *probably true*, etc. The total number of statements which a student judged accurately may be found by counting all of the tally marks in the cells labeled *a*, *g*, *m*, *s*, and *y*. This number may be expressed as a per cent of the maximum possible number of correct responses (150).

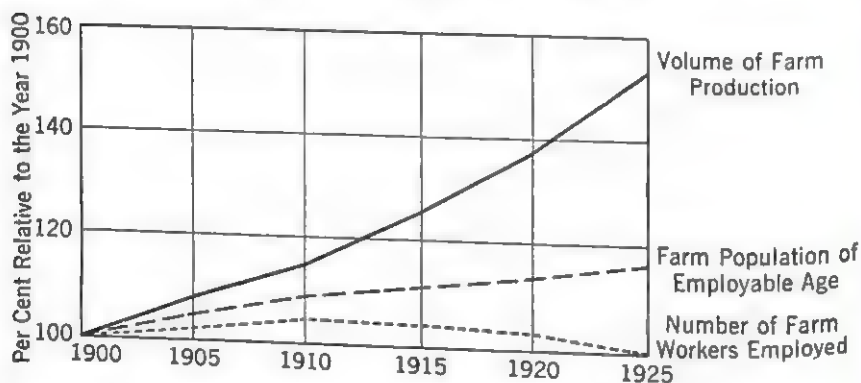
Since the judgment of the accuracy of the statements involves different levels of discrimination, depending on whether or not the interpretation needs to be qualified, it was found helpful to derive the following subscores on accuracy: (a) accuracy with *probably true* and *probably false*

⁵ In practice, the scoring may be done on the electric scoring machine, or if one is not available, by use of punched key stencils.

SAMPLE EXERCISE FROM FORM 2.52

- (1) are sufficient to make the statement true.
- (2) are sufficient to indicate that the statement is probably true.
- These Data Alone* (3) are not sufficient to indicate whether there is any degree of truth or falsity in the statement.
- (4) are sufficient to indicate that the statement is probably false.
- (5) are sufficient to make the statement false.

PROBLEM 1. This chart shows production, population, and employment on farms in the United States for each fifth year between 1900 and 1925.



Statements

1. The ratio of agricultural production to the number of farm workers increased every five years between 1900 and 1925.
2. The increase in agricultural production between 1910 and 1925 was due to more widespread use of farm machinery.

3. The average number of farm workers employed during the period 1920 to 1925 was higher than during the period 1915 to 1920.
4. The government should give relief to farm workers who are unemployed.
5. Between 1900 and 1925, the amount of fruit produced on farms in the United States increased about fifty per cent.
6. During the entire period between 1905 and 1925 there was an excess of farm population of employable age over the number of people needed to operate farms.
7. Wages paid farm workers in 1925 were low because there were more laborers than could be employed.
8. More workers were employed on farms in 1925 than in 1900.
9. Since 1900, there has been an increase in production per worker in manufacturing similar to the increase in agriculture.
10. Between 1900 and 1925, the volume of farm production increased over fifty per cent.
11. Farmers increased production after 1910 in order to take advantage of rapidly rising prices.
12. The average amount of farm production was higher in the period 1925 to 1930 than in the period 1920 to 1925.
13. Between 1900 and 1925, there was an increase in the farm population of employable age in the Middle West, the largest farming area in the United States.
14. Farm population of employable age was lower in 1930 than in 1900.
15. The production of wheat, the largest agricultural crop in the United States, was as great in 1915 as in 1925.

CHART SHOWING HOW SCORES ARE DERIVED

Student Responses \ Jury Key	Jury Key				
	True	Probably True	Insufficient Data	Probably False	False
True	Accurate a	Beyond Data b	Beyond Data c	Crude Error d	Crude Error e
Probably True	Caution f	Accurate g	Beyond Data h	Crude Error i	Crude Error j
Insufficient Data	Caution k	Caution l	Accurate m	Caution n	Caution o
Probably False	Crude Error p	Crude Error q	Beyond Data r	Accurate s	Caution t
False	Crude Error u	Crude Error v	Beyond Data w	Beyond Data x	Accurate y

statements, (b) accuracy with *insufficient data* statements, and (c) accuracy with *true* and *false* statements. They indicate the extent to which the student agrees with the jury in judging these three types of statements taken separately.

The first of these subscores may be computed by counting the tallies in cells *g* and *s*, and expressing this number as a per cent of the maximum possible number of such responses (59 in the case of the test under discussion). The second subscore mentioned above is derived from the number of tallies in cell *m* (expressed as a per cent of 61). The third

subscore is derived from the number of tallies in cells *a* and *y* (expressed as a per cent of 30).

The *going beyond the data* score indicates the extent to which the student marks statements keyed probably true as *true*, statements keyed insufficient data as *probably true* or *probably false*, and statements keyed probably false as *false*. The student is then granting the interpretation greater certainty than is warranted by the data.

In order to determine how frequently a student has "gone beyond the data," one may count the tallies in the cells labeled *b*, *c*, *h*, *r*, *w*, *x*. There are 120 opportunities for the student to react in this way, and the per cent of such responses may easily be calculated.

The *caution* score indicates the extent to which the student marks statements keyed true as *probably true*, statements keyed probably true as based upon *insufficient data*, statements keyed false as *probably false*, and statements keyed probably false as based upon *insufficient data*. The student is then refusing to attribute to the interpretations as much certainty as the jury was willing to do.

The *crude errors* score indicates the extent to which the student marks true or probably true statements as *false* or *probably false*, or marks false or probably false statements as *true* or *probably true*. This type of error is often due to carelessness in reading the data or interpretations, or to a misunderstanding of some terms involved in the data. Both of the last two scores may be computed in the manner prescribed for previous scores.

Omissions are scored in order to determine the actual number of opportunities the student had to score in other columns.

A form of data sheet on which scores from this test are conveniently summarized is presented on page 57. The scores made by seven students in the twelfth grade were selected for purposes of illustration. At the bottom of the

her class. In the case of the student called Homer, the pattern of scores indicates that he recognized the limits of the given data with an accuracy about equal to the average for his class. When he failed to judge accurately the limitations of the given data, Homer was overcautious in more judgments and went beyond the data in fewer judgments than was average for his class.

The *second* question that the test scores should answer is: How accurately does the student perceive various types of relationships in the data?

By examining the scores in columns 2, 3, 4, and 8, some tentative answers to this question may be obtained. As stated above, the score in column 1 gives the per cent of accuracy with which the student is able to judge limitations of interpretations dealing with all of the types of relationships in the test. Scores in columns 2, 3, and 4 are subscores of the *general accuracy* score. Each subscore refers to the accuracy with which the student judges certain of the relationships involved in the interpretation. For example, column 2 gives the per cent of accuracy of a student in recognizing those statements which are *probably true* or *probably false*. A high score here indicates that the student persistently applies with success the principles of prudent extrapolation, interpolation, and sampling. Column 3 gives the per cent of accuracy in judging statements which cannot be justified without the use of information from other sources. These statements include relationships such as cause, effect, purpose, analogy, as well as some statements of extrapolation, interpolation, and sampling. Column 4 gives the per cent of accuracy of a student in recognizing those statements which are *true* or *false*. A high score indicates that the student is able to judge accurately statements that involve comparisons of points in the data, and recognition or comparison of trends. The per cent of *crude errors* (column 8) indicates errors in which the student marked interpretations true that the jury considered

false or probably false, and vice versa. Such errors may be due to vocabulary or reading difficulties, carelessness, or inability to identify the relationship involved.

The following examples may help to clarify this explanation. Peggy's score in column 2 indicates that she stands low in relation to her group in the ability to make the finer discriminations necessary to judge accurately those extrapolation, interpolation, and sampling statements which are based on trends in the data. She is relatively poor in the accuracy with which she judges statements based on insufficient evidence, cause, effect, or purpose, as well as those extrapolation, interpolation, and sampling items that fall in this category. The score on accuracy with true and false statements (column 4) seems to indicate an ability approaching the average for her class in recognizing trends and comparisons of trends or of points in the data. However, this can be determined only after studying the entire pattern of scores.⁸ In view of Peggy's evident tendency to "go beyond the data," the higher score in column 4 may be a result of her tendency to be "gullible" and to mark many statements as true or false.

Homer's scores in columns 2, 3, and 4 seem to indicate a greater accuracy in his judgment of statements based on insufficient data than on the statements classified in the other two categories. However, it is necessary again to consider the entire pattern of scores to make a justifiable inference. Homer's relatively high score on *caution* and low score on *beyond data* imply that he tends to refuse to make judgments

⁸ Intercorrelations have been computed to investigate the extent to which scores described above are statistically independent. See Appendix. Although positive correlation exists between each of the subscores on general accuracy, the intercorrelation is not sufficiently high to permit the prediction of one score from another. However, a high negative correlation exists between the score on *beyond data* and *insufficient data*, and between *general accuracy* and *crude errors*. From a statistical standpoint it is possible in both these cases to predict one of these scores from the other without appreciable loss of information about the student, but teachers find it less difficult to interpret the individual scores when all these scores are retained.

of probability and classifies statements that are not well justified by the data as of the *insufficient data* type.

An examination of scores made by Joseph and Andrew shows that, although both boys receive the same score in *general accuracy*, for those judgments in which they fail to be accurate Andrew tends to go beyond the data more often than Joseph.

It is usually inadvisable to interpret scores on this test in terms of national norms, since opportunities to develop these abilities vary markedly from group to group. Data on means and standard deviations for certain groups are given in tables in the Appendix. If a group is known to be comparable to these groups, these statistics may be helpful as a background of comparison.

OTHER INSTRUMENTS TO MEASURE THIS OBJECTIVE

During the period of the Eight-Year Study a number of instruments were developed for exploration of the ability to interpret data. Responses on some of these were useful in pointing out a need for further clarification of the objective. Statistical studies of responses led to changes which were incorporated in subsequent forms. In some forms of the test, modifications were introduced to meet the particular needs of different teachers. The purpose of the discussion that follows is to give a brief survey of the changes that have taken place in the test and the reasons for them.

One of the earliest tests that explored certain aspects of this objective was constructed to measure "the ability to infer."⁹ One short-answer form of this test required the students to judge the best of five given inferences. A study of the responses on this test and a corresponding essay form yielded many clues concerning the types of inferences that students make. A higher validity coefficient was secured

⁹ R. W. Tyler, "Measuring the Ability to Infer," *Educational Research Bulletin*, IX (Nov. 19, 1930), p. 475.

when the students were required to judge both best and worst inferences than when they judged only the best.

Results of exploratory tests using a three-response form and others using a five-response form yielded valuable information concerning the objective. In one of the earliest of the five-response forms of the test, the student was presented with different types of data and asked to judge interpretations made from them. The directions were as follows:

Consider carefully each of the following statements, and indicate in the columns to the right whether you believe:

1. *the data alone* justify the statement.
2. *the data alone do not* justify the statement.
3. *the data together with your information* suggest that the statement is probably true.
4. *the data together with your information* suggest that the statement is probably false.
5. *the data together with your information* are insufficient to make a decision concerning the statement.

This form was used in an attempt to get evidence of two kinds of behavior in interpretation of data, namely, ability to adhere rigidly to the data and reject interpretations that go beyond or are contradicted by the data; and the ability to draw meaningful inferences from those interpretations which go beyond the data but which appear highly probable or improbable in the light of other information known to students. Difficulty was encountered in interpreting these scores, since there was no way of setting up controls or standards for judging the amount or quality of outside information a student was using in judging the inferences presented. As will be recalled, the definition of the objective accepted by the committee emphasizes the ability of the students to recognize what the given data reveal, and to distinguish acceptable inferences from those that cannot be justified without using information or principles from other sources. This

restriction led to a reformulation of the directions, and thereafter they remained virtually the same in subsequent forms of the test.

Teachers of several subject fields were interested in this objective. To meet their request some of the first forms in which the revised directions were used restricted the field from which the data were drawn to the natural sciences or the social sciences.¹⁰ Since it was believed that the behaviors involved in these forms are not essentially different, it was deemed advisable to reduce the time required in measuring this objective by measuring in one instrument the achievement relative to several fields. Thus subsequent forms included in the same booklet data drawn from both fields.¹¹ Statistical considerations (e.g., studies of reliability) indicate that this has not changed the homogeneity of the behavior to any great extent.

The summarization of scores has remained, with one exception, very much as it is found on the sample data sheet given above for Form 2.52. In early forms (2.2, 2.3, 2.4) the *beyond data* scores had subscores which indicated the tendency of the student to go beyond the data in the direction of greater truth or in the direction of greater falsity than the data warranted. From an analysis of responses it was found that in general most students showed much greater tendency to go beyond the data in the direction of judging the printed statement as true than in judging it as false. Because of this fact these subscores on "going beyond the data" did not greatly aid the interpretation of scores and were omitted from subsequent forms of the test. A *caution* score that was found to be more meaningful in describing the behavior of students was added.

A statistical study of student responses to Form 2.5 sug-

¹⁰ Form 2.2, Interpretation of Data (Natural Sciences) and Form 2.3, Interpretation of Data (Social Sciences).

¹¹ Forms 2.4, 2.5, 2.51, 2.52, Interpretation of Data.

gested that greater reliability of certain scores could be obtained by increasing the number of statements of each type used in the test. These suggestions were used in building Form 2.51 by including in each of the ten exercises 15 statements which constituted a definite pattern of types of interpretations and types of responses expected. An effort was made to include in each exercise at least one statement involving each type of relationship used in the test, but statements including extrapolation, interpolation, and sampling were used in greater number. The entire test was thus lengthened from 119 statements in Form 2.5 to 150 statements in Form 2.51 and the *probably true* or *probably false* response was expected in 40 per cent of the statements.

The latest form of Interpretation of Data test (Form 2.52) was intended to be comparable to Form 2.51. An effort was made to match the form of presentation, types of interpretations, topics with which the data deal, and types of response expected. Each of the two forms was administered within a week to 105 students of the tenth grade, 133 students of the eleventh grade, and 99 students of the twelfth grade of two large high schools. The coefficient of correlation between the two forms of the test for each category was computed by the product-moment method. These coefficients, together with means and standard deviations on each category for both tests, are given in Table 1 below.

Although these correlations are fairly high, the fact that they are no higher may be partially explained by the observation that more rigorous standards were used in keying Form 2.52 and that some sources of ambiguity found to be present in Form 2.51 were eliminated.

Since some teachers were interested in measuring the abilities of junior high school students in interpreting data, a form was developed for students of this grade level. The criteria for the selection of data were similar to those used in Form 2.52, and the advice of junior high school teachers

and librarians was sought in checking the appropriateness of the data and the interpretations for students of this level. As a result of this advice, an attempt was made to simplify this

TABLE 1

Means and Standard Deviations for Tests 2.51 and 2.52; Product-Moment Correlations between Forms 2.51 and 2.52.

Category		General Accuracy	PT PF	Insuf. Data	TF	Cau-tion	Beyond Data	Crude Er-rors
Means Standard Deviations	2.51	40.1	26.5	38.6	52.3	27.8	51.4	16.9
	2.52	45.2	24.4	53.7	70.6	26.4	37.5	13.8
	2.51	10.4	14.0	15.2	16.2	11.0	12.4	5.97
	2.52	11.4	14.4	18.4	16.2	12.7	13.4	6.30
$r_{2.51, 2.52}$.85	.84	.83	.74	.85	.81	.65

instrument, in comparison with Form 2.52, in vocabulary, in types of responses expected, in number of interpretations used, and in problem areas or concepts involved. A preliminary form (2.7) was constructed and administered, and after a statistical study of the responses, the suggested improvements were incorporated in the present test, Form 2.71.

This test contains ten sets of data, each of which is followed by ten interpretations. The data deal with problems of safety, budgeting, sports, choice of vocation, cost of living, etc. The student is required to make three distinctions in judging these interpretations. These are given in the directions of the test as follows:

- A. Enough information is given to make the statement *true*.
- B. Not enough information is given to decide.
- C. Enough information is given to make the statement *false*.

The student responses are summarized in terms of scores briefly denoted by the following phrases: general accuracy,

caution, beyond data, crude errors, accuracy with true—false, and accuracy with insufficient data. Reliability coefficients were computed by the Kuder-Richardson formula for five populations drawn from each of grades seven, eight, and nine.¹² For these 15 populations the reliability coefficients of the *beyond data* and *insufficient data* scores are of the same order of magnitude as are those of the *general accuracy* score. The reliability of the other scores analogous to those of Form 2.52 are a little lower with the exception of those for *crude errors* which, as one might expect, are erratic and tend to be rather low. This same general pattern is found for each grade.

VALIDITY OF THE INTERPRETATION OF DATA TESTS

Two main aspects of the validity of the interpretation of data tests will be considered: (1) the validity of the tests as a measure of the students' ability to judge interpretations formulated by others, and (2) the validity of the tests as an index of students' ability to write original interpretations.

Ability to Judge Interpretations Made by Others

The validity of this test as a measure of the ability to judge interpretations made by others depends upon several factors: (a) the correspondence between the behaviors demanded of students in the test and the behaviors defined in the statement of the objective, (b) the adequacy of sampling relative to form of presentation, to problem areas with which the data are associated, and to types of interpretations, (c) the appropriateness of the test as to difficulty for the high school level.

In considering the first point, it should be recalled that the test is so constructed as to afford the student an opportunity

¹² G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika*, Vol. 2, No. 3 (Sept., 1937), pp. 151-160. Throughout this report, wherever the Kuder-Richardson Method is indicated, case III of this method was used. These and other data on Form 2.71 will be found in the Appendix.

to demonstrate the two main behaviors defined in the objective, namely, the ability to perceive relationships in the data and the ability to recognize the limitations of the data. To verify this, it will be necessary to review briefly the method of construction of the test. Incorporated in the interpretations which the student is asked to judge are the various types of relationships, such as trends, comparisons, etc., that he is expected to perceive, expressed in such a way as to have varying degrees of substantiation from the given data. Thus some statements are intended to be fully established or contradicted by the data alone, some statements if properly qualified are partially established or contradicted by the data, and others are unjustified without the use of information from other sources. The five-point response by which the student indicates his judgment of the interpretations forces a response by the student from which the extent of his recognition of the limitations of the data and his perception of relationships may be inferred.

It should also be recalled that the criteria for selection of data were determined by the judgment of members of the committee. Their knowledge of types of materials that students read and an analysis of the types of data commonly found in curricular and other reading materials form the basis of their judgment of the adequacy of the sampling of forms, of presentation, of problem areas, and of types of interpretations. The analysis made by E. W. Hellmich of textbooks for social studies in the junior and senior high school and in elementary college courses indicates that the subject matter and types of presentation of the data used in Test 2.52 are those which students encounter.¹³

¹³ Eugene W. Hellmich, *Mathematics in Certain Elementary Social Studies in Secondary Schools and Colleges*, Teachers College, Columbia University, Contributions to Education, No. 706, 1937. Studies in other fields report similar results: for example, Robert C. Scarf, *Mathematics Necessary for the Reading of Popular Science*, Master's Thesis, The University of Chicago, Department of Education, 1925.

The appropriateness of the test for the high school level can be considered in terms of two sources of evidence. First, the interpretations represented by the statements in the test are of the types students are found to use when they make their own interpretations of data. Secondly, study of the distribution of scores made by students who have taken the test shows that no student from the ninth grade to the junior college level has received the maximum score possible, nor is there concentration of scores at the lower end of the range. The distribution of scores is symmetrical with concentration of scores at the mean, and, in general, the means tend to increase with grade level.

Ability to Make Original Interpretations

Although teachers are interested in appraising students' ability to judge interpretations made by others, many teachers wish also to measure the students' ability to make their own interpretations. In order to use scores on the interpretation of data test as an index of the latter ability, there must be evidence of a high correlation between scores on the test and judgments of the students' ability to make original interpretations. To obtain such evidence, attempts were made in earlier studies to validate the interpretation of data test by using free essay responses of students as a criterion. For example, in a study conducted in a large public junior high school in which 193 students of seventh, eighth, and ninth grades participated, the students were given the sets of data taken from an Interpretation of Data test for the junior high school level (Form 2.71) and were asked to make free essay responses following such general directions as: "Write five statements that you are sure are true according to the facts given in these data," and "Write three statements based on the data which you are not quite sure are true according to these data." *

The objectivity secured in grading this essay form is indi-

cated by the values of the product-moment coefficients of correlation among the three judges who marked the papers. These values ranged from 0.92 to 0.96. Table 2 below gives the values of the product-moment coefficient of correlation between Form 2.71 and the essay form, and the reliabilities of each form of the test.

TABLE 2
Statistics for General Accuracy Score of Test 2.71

Grade	N	Product-Moment Correlation between Test 2.71 and Essay Form	Reliability Coeffi- cient of Essay Form by Split- Halves Method with Spearman- Brown Correction	Reliability Coeffi- cient of Test 2.71 by Kuder- Richardson Method
7	68	0.69	0.88	0.80
8	60	0.58	0.73	0.87
9	65	0.44	0.79	0.91

The correlations between the two forms were positive and sufficiently large to warrant a further investigation of the relationship between the behaviors involved.

Although a wide range of relationships, such as comparisons and recognition of trends, was found in the statements made by students, as a rule the free responses made by any one student involved a narrow range of relationships, and did not sample adequately his ability to make various types of interpretations. In the next study, directions on the essay form of the test were changed in an effort to encourage the student to include a wider range of relationships in his interpretations. The new directions posed a series of questions designed to direct the attention of the student to the various types of relationships found in the interpretations given in Form 2.52. For example, after each of the following inter-

pretations is the question which corresponded to it in the essay form:

- 1a. The ratio of agricultural production to the number of farm workers increased every five years between 1900 and 1925. (Comparison of trends)
- 1b. In terms of these data alone, what do you believe you can say concerning (a) the change in number of farm workers employed compared to (b) the change in volume of farm production throughout the period recorded in the chart?
- 2a. The increase in agricultural production between 1910 and 1925 was due to more widespread use of farm machinery. (Cause)
- 2b. In terms of these data alone, what do you believe you can say about the cause of the increase in volume of farm production between 1910 and 1925?
- 3a. The average amount of farm production was higher in the period 1925 to 1930 than in the period 1920 to 1925. (Extrapolation)
- 3b. In terms of these data alone, what do you believe you can say about the volume of farm production during the period from 1925 to 1930?

This study was made with two populations of ninth, tenth, eleventh, and twelfth grade students. One group consisted of 119 students from a large public high school and the other was made up of 99 students from a smaller private high school. The essay form was administered first, followed within a week by the regular form of Form 2.52.

The essay responses were scored and summarized so that statements involving each type of relationship could be classified as *accurate*, *beyond the data*, *cautious*, *involving a crude error*, or *unable to see the relationship*. In scoring, it was possible by the use of a simple set of rules to score papers so objectively that correlations of the scores given inde-

pendently by three markers ranged from .94 to .96. The additional time required to answer the essay form made it necessary to sample the types of relationships and the types of data used in Form 2.52. Seven questions were formulated for each of six of the ten exercises in Form 2.52; each of the 42 questions thus formulated corresponded in subject matter and type of relationship to a statement used in that test. Only 39 answers were scored in the essay form because three questions were later found to be ambiguous. These were a fair sample of the whole test, since a product-moment correlation coefficient of .85 (uncorrected for overlapping) was obtained between the "general accuracy" score on these 39 items and on the entire 150 items of Form 2.52. Since the correlation between the part and the total test was desired as a measure of the adequacy of the sampling, no correction for overlapping was made. There was also a product-moment correlation coefficient of .96 between the *general accuracy* scores of the entire ten exercises of Form 2.52 and the six exercises from which these 39 items were taken. However, there does appear to be some difference in the difficulty of the 39 items and of the total test. The mean general accuracy score for the 39 items was definitely higher than that for the total 150 items for each of the two different populations of approximately 100 high school students. In spite of this difference, however, the sample appeared to be sufficiently representative for use in this validity study.

The scores on the essay form were correlated by the product-moment method with scores on similar categories for Form 2.52. The results are given in Table 3 below.

The reliabilities of the essay form for these populations were computed by the Kuder-Richardson formula and are found in Table 4. Reliabilities for Form 2.52 will be found in Table 5 under the discussion of reliability.

Since the correlation coefficient is to be used as a measure

TABLE 3

Correlations for Each Category between Essay Form and Form 2.52

Correlations for Each Group															
Score	General Accuracy			Beyond Data		Caution		Crude Error		True-False		Insufficient Data		Probably True Probably False	
	N	r	r _{corr.}	r	r _{corr.}	r	r _{corr.}	r	r _{corr.}	r	r _{corr.}	r	r _{corr.}	r	r _{corr.}
Small Private School	99	.72	.80	.60	.65	.50	.55	.22	.56	.37	.47	.64	.71	.53	.63
Large Public School	119	.74	.83	.47	.52	.51	.57	.08	.12	.58	.77	.58	.65	.55	.66

r_{corr.} refers to reliability coefficient corrected for attenuation due to the unreliability of the criterion.

TABLE 4

Reliabilities by Kuder-Richardson Formula for Two Populations on Essay Form

Score	General Accuracy	Beyond Data	Caution	Crude Error	True-False	Insufficient Data	Probably True-Probably False
Small Private School	.81	.85	.82	.15	.61	.82	.70
Large Public School	.80	.82	.81	.43	.57	.80	.70

of validity (that is, of the degree to which the ability to make original interpretations of data can be predicted from a score on Form 2.52), it does not seem legitimate to correct for attenuation due to the unreliability of Form 2.52. The relation between the theoretical ability to judge interpretation and the theoretical ability to make original interpretations is not at issue, but rather how well Form 2.52 predicts the latter ability. Hence, it seems defensible to correct for

the unreliability of the criterion but not for that of Form 2.52. As seen in Table 3, such correction yielded validity coefficients of .80 and .83 for the *general accuracy* score, and lower values for the other categories. A validity coefficient of .80 is sufficiently high for group predictions and is of some value for study of individual students. Thus Form 2.52 can be used as an index of the general accuracy with which a group can make original interpretations of data. For the populations used in this study, its validity as an index of the types of errors into which students fall in making original interpretations was not high.

Some differences in the two forms of the test are apparent. In the essay form the student could respond with more than one statement or could make an irrelevant statement—that is, a statement in which he failed to involve the relationship to which the question was intended to direct his attention. There was no opportunity in Form 2.52 to react in either of these ways. However, since the relevant responses to each question on the essay form were scored as a whole on the basis of the main thought expressed, the number of extra statements did not affect the score. The irrelevant statements affected the score on general accuracy in the same way that an omitted item would have affected this score on either form. A study was made to determine whether the opportunity in the essay form to respond with irrelevant statements might be an important factor affecting the correlation between the two instruments. The correlation coefficient between the general accuracy score of the essay form and all of the corresponding 39 items of Form 2.52 for the group of 99 students was .68. A general accuracy score on Form 2.52 was derived for only that part of the 39 items to which the student had made *relevant* responses on the essay form. The product-moment correlation coefficient between the general accuracy score on the essay form and this part score was found to be

.78. This seems to indicate that the opportunity to make irrelevant responses on the essay form may be one of the factors that limits the correlation.

The comparison of *patterns* of responses for the same individuals on the two test forms suggests another likely hypothesis to account for the differences in results. Many students apparently employed somewhat different standards in making original interpretations than they used when judging interpretations of data made by others. Students' behavior in this respect may be classified into the following patterns:

- a. The student reacts similarly on corresponding items of the two forms.
- b. The student is overcautious on an item in judging interpretations made by others but goes beyond the data on the corresponding item in making his own interpretations. The reverse pattern also appears.
- c. The student is either very cautious or goes beyond the data in judging interpretations made by others, but is accurate when making his own interpretations. Here also the reverse pattern appears.

Of these patterns, the first appeared most frequently, as might be expected from the high validity coefficients. Extreme discrepancies between reactions on corresponding items of the two tests (as described in pattern b) appeared very infrequently. In pattern c, students tend to go beyond the data more in making their own interpretations of data than in judging interpretations made by others.

While other factors may be present, the differences between the essay form and Form 2.52 may in part be attributed to the opportunity in the essay form to make irrelevant statements, and to the tendency of some students to use different standards in reacting to corresponding items of the two forms.

RELIABILITY OF THE INTERPRETATION OF DATA TESTS

The most comprehensive study of reliability of Form 2.52 was made by the use of the Kuder-Richardson formula with 19 populations from grades nine, ten, eleven, and twelve in seven schools. The reliabilities for the two populations used in the validity study were of special interest and are given in Table 5 below. The means and standard deviations for these two populations are listed in Table 6 below.

TABLE 5
*Reliabilities by Kuder-Richardson Formula
on Form 2.52 for Two Populations*

Score	N	General Accuracy	Beyond Data	Caution	Crude Error	True-False	Insufficient Data	Probably True—Probably False
Small Private School Grades 9, 10, 11, 12	99	0.93	0.91	0.91	0.75	0.78	0.92	0.88
Large Public School Grades 9, 10, 11, 12	119	0.95	0.93	0.87	0.81	0.84	0.90	0.88

It will be noted that the reliability coefficients in all categories except *crude error* and *true-false* cluster around .90 for both of these populations and that the general accuracy score has the highest reliability. The coefficients tend to form the same definite pattern from category to category for both populations, and the difference between the coefficients for the two populations on any single category is slight.

TABLE 6

*Means and Standard Deviations of Per Cent Scores
on Form 2.52 for Two Populations*

Score	General Accuracy			Beyond Data		Caution		Crude Error		True-False		Insufficient Data		Probably True—Probably False	
	N	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ	M	σ
Small Private School	99	56.3	10.9	19.6	11.2	36.1	13.5	7.8	5.3	78.3	15.0	76.8	16.7	24.3	14.1
Large Public School	119	45.9	13.7	47.6	13.8	24.5	10.3	13.2	7.0	62.0	17.3	41.3	17.5	34.1	16.1

When the means and standard deviations for the two samples are considered, it will be noticed that the group from the small private school is in general a superior group as measured by Form 2.52. It is also a more cautious group as measured by the high mean score on *caution* and by the low mean score on accuracy with *probably true—probably false*. Yet in spite of the difference in these two groups, the reliabilities computed from them are very similar. Table 1 in the Appendix gives the reliability coefficients for all nineteen populations. It will be noted again that for these populations the reliability coefficients of all scores except *crude errors* and *accuracy with true and false statements* are sufficiently high for group interpretation.

Before Form 2.52 was made, the split-half method was used in deriving the reliability of Form 2.51. An effort was made to split the test into "equivalent" halves by pairing items according to definite criteria, such as the response expected of the student, the types of interpretation involved, the topic with which the data dealt, and the form of presentation of the data. An analysis of the responses of 88 students was used in an attempt to include in each half items which presented these students with the same type of difficulty, but

it was not always possible to make an accurate match. The correlation between "equivalent" halves of Form 2.51 was computed from the scores of another population of 284 students in the three upper grades of two high schools. By means of the Spearman-Brown formula it was possible to predict the correlation for a test doubled in length. Table 7 contains these corrected correlations.

The coefficients obtained from the comparability study discussed previously may be considered another measure of reliability of the interpretation of data test and are also given in Table 7 below. However, the lower values of these coefficients are attributable more to the difference between the two tests than to the unreliability of either of the tests.

TABLE 7
Reliability Coefficients for Interpretation of Data Tests

Method	Population	N	General Accuracy	Beyond Data	Caution	Crude Error	True-False	Insufficient Data	Probably True—Probably False
Kuder-Richardson Form 2.52	Grades 9, 10, 11, 12	119	0.95	0.93	0.87	0.81	0.84	0.90	0.88
Comparability Forms 2.51-2.52	Grades 10, 11, 12	337	0.85	0.81	0.85	0.65	0.74	0.83	0.84
Split-halves Form 2.51	Grades 10, 11, 12	284	0.92	0.91	0.91	0.82	0.86	0.92	0.87

When the reliabilities obtained by the three methods are compared, it will be noted that the coefficients computed by the Kuder-Richardson formula and by the split-halves method are approximately the same and that, as would be expected, the coefficients computed from scores on "comparable" forms are smaller for all categories. These reliabilities were considered rather high in view of the complexity of the behaviors involved.

II. APPLICATION OF PRINCIPLES OF SCIENCE

ANALYSIS OF THE OBJECTIVE

Teachers of science in schools of the Study believed that students should learn to apply knowledge obtained in the science classroom and laboratory to the solution of problems as they arise in daily living. This aspect of critical thinking was frequently mentioned in the list of objectives submitted to the Evaluation Staff. A study of the prevailing curriculum materials for science instruction confirmed the importance of this objective, and therefore a committee was formed for the purpose of clarifying it and of aiding in the development of evaluation instruments for appraising growth in the ability to apply science information. Although this objective had previously been explored to some extent at the college level by Tyler¹⁴ and others, and these explorations had served to show that certain techniques for the measurement of the objective were feasible, it could not be assumed that the available analyses and methods were immediately applicable at the secondary school level. This committee of teachers in the schools therefore aided the Evaluation Staff in clarifying the objective to be appraised and also in finding situations which would give students an opportunity to show the degree to which the objective had been attained. In the present instance, clarifying the objective necessitated an analysis of the behaviors involved in *application* and a selection of the principles to be used.

Behaviors Involved in Application

The analysis of the behaviors involved in *application* separated the process of applying principles into two steps: (1) the student studies a situation and makes a decision about the probable explanation or prediction which is ap-

¹⁴ Ralph W. Tyler, *Constructing Achievement Tests*, Bureau of Educational Research, Ohio State University.

plicable to this situation; (2) he justifies through the use of science principles and sound reasoning the explanation or prediction that he made in the first step. In the first step he acts in the role of an authority who is presented with a problem and asked for a solution. In the second step, he is asked to explain or justify that proposed solution by means of his previous knowledge of what has occurred in similar situations.

The kind of deductive thinking needed for the solution of these problems consists of the search for an explanation of the fact or facts described in the problem situation by means of some *general rule* which asserts a highly probable connection between facts of the kind described in the problem and other facts the student knows to be applicable to similar problems. The question he attempts to answer is: Does the general rule which is suggested by the given facts as an hypothesis for explaining what has happened (or what will happen) actually apply to this specific problem? The answer to this question comes, of course, from experimentation or direct observation. However, if observations have been made in several situations which can be shown to be similar to that one which is described in the test, then without obtaining the empirical evidence one may nevertheless predict with considerable confidence that the same conclusion is also true in this case. It was for the measurement of such behavior that the instruments to be described later were constructed. The teachers felt they needed the most help in evaluating the ability of students to apply principles in *new* situations, and consequently the *remembering* of applications which had been made was not included as a behavior to be directly appraised.

Selection of the Principles

In the discussions that were held to clarify the meaning of the term *principle* it was found that some teachers were

inclined to accept certain statements as representing "principles" whereas others wanted to regard them as statements of "facts." The difficulty was resolved by obtaining an agreement which permitted, for the purpose of testing application, the use of any science information, fact, generalization, understanding, concept, or "law" which proves to be useful (alone or in connection with other information) for predictive or explanatory purposes. Although more inclusive than the definition of principle that is frequently used by science teachers, this agreement seemed satisfactory for the measurement of the objective as this committee conceived it. After the committee had accepted this agreement as to the "principles" which were to be used in the construction of the instruments, teachers were asked to submit statements of those principles which were considered important in their courses and which had received the greatest emphasis in their teaching. These lists included the principles with which their students had had the greatest opportunity to become familiar through reading, discussion, and experimentation.

The original lists from individual teachers included principles from the fields of chemistry, physics, and biology, as well as some that were common to all three fields. After the principles submitted had been classified into subject-matter areas, the complete list was sent to a number of teachers in the Thirty Schools. These teachers were asked to:

1. Select those statements that they would expect their students to apply in making predictions or explanations in new situations.
2. Select those statements that they would expect their students to know in a general way, but not to the extent of being able to use them to make predictions in new situations.

Only those principles which were included in the first category by at least three-fourths of the teachers were con-

sidered for use in the tests. Two additional criteria were established to aid in the selection:

3. The principle should have a wide range of applicability to commonly occurring natural phenomena.
4. The principle, with examples of its application to commonly occurring phenomena, should be found in all of the science textbooks commonly used in these schools.

The teachers were also asked to judge the relevance of each principle to the areas of general science, biology, chemistry, or physics, or to all of these areas.

THE DEVELOPMENT OF EVALUATION INSTRUMENTS

During the period of the Eight-Year Study a number of instruments were developed for evaluating the ability to apply principles. Several of these instruments included principles drawn from the subject-matter area of general science; others were restricted to principles drawn from physics, chemistry, or biology. Because the instruments which included principles from general science were used more extensively than the others and because they were the ones experimented with in attempting to arrive at a satisfactory pattern for the test, they will be used to illustrate the construction of tests of application of principles.

Preliminary Investigations

In preparing a test of Application of Principles, the first step after the principles had been selected was to obtain problem situations to which the student might react. Teachers were asked to submit to the committee problem situations which:

1. were new to the students (i.e., they were not ordinarily discussed in the classroom or used in the textbooks);

2. occur rather commonly in actual life;
3. could be explained by the principles which the teachers had selected as important for their students to apply.

Attempts to phrase the problem situations revealed that they might be so described as to demand several different types of response from the student. Four types of response were used; namely, making a prediction, offering an explanation for an observed phenomenon, choosing a course of action, and criticizing a prediction or explanation made by others. An illustrative situation of each type follows:

1. *A farmer grafted a Jonathan apple twig on a small Baldwin apple tree from which he had first removed all the branches. The graft was successful. If a new branch develops from a bud below the point of the graft and produces apples, what kind of apple will it be?* Here the student is asked to make a prediction about a situation in which presumably he has had no actual experience. It is presumed that if he understands certain laws of heredity, he will be able to make a valid prediction.
2. *All of the leaves of a growing green plant were observed to be facing in the same direction. Under what conditions of lighting was the plant probably grown?* This example requires that the student offer an explanation of an observed phenomenon. Some knowledge of the principles of photosynthesis, growth, and tropistic responses of plants would be required for the solution of this problem.
3. *The rear of an automobile on a wet pavement is skidding toward a ditch. If you were the driver of the car, what would you do to bring the car out of the skid?* This problem requires the student to choose a course of action. A knowledge of the principles of

centrifugal force and Newton's laws of motion would enable the student to choose a satisfactory *course of action*.

4. *It was reported in a newspaper that in order to tow down a river a huge oil drum filled with air, the workmen found it necessary to fill the drum with compressed air to increase its buoyancy. Do you believe that this would increase the buoyancy of the oil drum?* This problem asks the student to criticize an explanation which has been given. Knowledge of the fact that air has weight and of the principles of buoyancy are required for a satisfactory solution in this problem.

In none of these problems were the answers expected to be in exact quantitative terms; rather a qualitative understanding of the general outcome was required. It was thought that the kind of activity shown by students in making a prediction of this kind was of more importance for general education than one which required exact substitutions of numerical data in a formula or similar activities frequently used in the laboratory. One often encounters problems in which a principle is used to explain what happens in general when certain factors are varied in the situation, while the need for numerical solutions of problems occurs relatively infrequently for most people. Although the above problem situations are stated in such a way that the student is expected to react somewhat differently in each, it is not likely that he will react intelligently to any of these situations unless he has a knowledge of the principles operating and has recognized their application to the problem. Whether he criticizes a prediction made by someone else or makes the prediction himself, he must base his answer upon the knowledge which he feels is applicable to the situation.

The next step in constructing the test was to determine the

reasons which might justify the response to the problem situation, and to find a means of appraising the reasons cited by the student. Science teachers were in rather general agreement that the most valid of all the reasons a student might use for justifying his conclusions would be those that cited established scientific facts, principles, and generalizations. However, in addition to these, it was agreed that the student might cite from his experience, from authoritative materials he had read, or he might use analogous situations familiar to the person to whom he was explaining his decision, provided these experiences, authorities, or analogies were pertinent to the situation he was attempting to explain.

In order to determine whether or not students did use these kinds of reasons, they were asked to write out both their own predictions, choice of action or responses to the situation, and all of the reasons that they believed would support the decision they had made. When these papers were analyzed by the teachers and the Evaluation Staff, the types of acceptable reasons which had been anticipated were found in the students' responses. However, in addition to the reasons which were agreed upon as being acceptable, certain types of errors were also found to occur rather consistently in the written responses of the students. It was found that students frequently used teleological explanations and analogies not closely correspondent to the situation described in the problem. They cited authorities that were questionable, ridiculed positions other than their own, stated as facts certain misconceptions or superstitions, merely restated either the facts given or their own prediction, and made less frequently a variety of other types of errors. They also used, in addition to the principles and facts judged to be acceptable and necessary to the explanation of the problem, other facts and principles that were irrelevant to the solution of the problem. The frequency with which each of these types of reasons was used was not constant, but varied from class

to class and from problem to problem. In examinations of sufficient length given to a large number of students, however, these types of errors were found to be most prevalent.

In general, it was possible to infer that the errors were made because:

1. The student did not know the principles.
2. He did not see that a principle he knew applied to the situation.
3. He knew the principle and knew that it applied to the situation, but he was unable to explain adroitly how or why it applied.
4. He used teleology, poor analogy, or poor authority, rather than (or in addition to) correct facts and principles.
5. Although his explanation was correct as far as it was given, he cited facts and principles which were inadequate for a convincing proof for a given selected conclusion or course of action.
6. He confused closely related principles, only one of which was applicable to the problem.
7. He used irrelevant material.
8. He neglected to study the description of the situation carefully enough to note all of the limiting factors in the description.

This list does not include all of the reasons why students made errors but it does help to show why it was difficult to score the written responses.

Construction of Early Short-Answer Forms

The same problems of objectivity of scoring and of adequate sampling that are found in any essay-type test were inherent in these written responses. The teachers found that it was difficult to differentiate among those acceptable uses of generalizations, facts and principles which were relevant

to the problem, and the logical errors, obscured as they sometimes were by illegibility of handwriting and by awkward literary style. It was also difficult to decide when a student had cited enough evidence to support his choice of answer. A second criticism of this form of test was that it limited the number of principles which could be sampled because of the time required by the student to write out the answers. Because of these difficulties, a more objective means of testing this same ability was sought.

Following a study of the responses written out by students, the first of a series of objective test forms in this area was made. The objective form of the test asked the student to select from a list of predictions for each problem situation the one which he thought was most likely to be true, and then to select from a list of reasons those which would be necessary to establish the validity of his choice. The predictions and reasons used in the test paralleled those which had been used frequently by the students when they wrote essay-type responses. When experimental groups were given an examination which required them to *write out* their predictions and reasons for the first half of the testing period, and an examination in which they were required to *select* the correct prediction and the reasons which supported it from a given list during the latter half of the period, it was found that the results on the two types of examinations were quite similar. The coefficient of correlation was in all cases above 0.80.¹⁵ The advantages of more objective scoring and the possibilities for more extensive sampling of problem situations led to the adoption of the objective form.

¹⁵ Ralph W. Tyler, *Constructing Achievement Tests*, Bureau of Educational Research, Ohio State University; Fred P. Frutchey, "Evaluating Chemistry Instruction," *Educational Research Bulletin*, XVI (Jan. 13, 1937); Louis E. Rath, "Techniques of Test Construction," *Educational Research Bulletin*, XVII (April 13, 1938); Louis M. Heil, "Evaluation of Student Achievement in the Physical Sciences—The Application of Laws and Principles," *The American Physics Teacher*, VI (April, 1938).

The procedures used in preparing the early form of objective tests in this area were as follows:

1. The principles to be used in the test were selected in accordance with the criteria formulated by the teachers interested in this objective.
2. Problem situations in which these selected principles would apply were chosen with the following criteria in mind:
 - 2.1 They were to be new in the sense that they had not been used in the classroom or laboratory.
 - 2.2 The situation should approximate a rather commonly occurring life situation.
 - 2.3 The problem should be significant to students in that its solution might help them to solve similar problems which occur in their everyday living.
 - 2.4 The vocabulary used should be at an appropriate level for the students taking the test. They should be able to understand the description of the situation.
3. Several (usually three or more) plausible answers for the problem were formulated. These might be in the form of predictions, courses of action to be taken, causes to be stated, or an evaluation of one of these when it was given. Actually, when possible answers were suggested by listing them in the test, the procedure in every case would be one of evaluation through the selection of what the student thought was the most desirable, whether it was a prediction, course of action or explanation for the phenomena which had been described in the problem.
4. Finally, reasons of the sort used by students were listed, including for each situation those common types of errors which students made when they wrote out their reasons. In addition to correct statements

of scientific principles needed for a satisfactory explanation, the following types of statements were formulated:

- 4.1 *False statements purporting to be facts or principles.* These, if accepted as true, would support one of the alternative conclusions. For example, if the correct principle stated that a direct relationship existed between two phenomena, one might word a false statement in such a way as to indicate that there was no relationship or that the relationship was an inverse one. To remain consistent in his reasoning, the student can use such a statement only to support a conclusion other than the acceptable one.
- 4.2 *Irrelevant reasons.* These statements are true, but either they have no relationship to the phenomenon described in the problem or they are quite unnecessary in the explanation of the phenomenon.
- 4.3 *False analogies.* These stated directly or inferred that the phenomenon described in the problem was identical with, or very much like, some other known phenomenon when it actually had little or nothing in common with it; therefore, an explanation for one phenomenon would not be acceptable for explaining the other. Metaphors were sometimes included as an example of a more subtle use of analogy, in that the analogy was implied by the use of words but not definitely expressed.
- 4.4 *Popular misconceptions.* These included the more common beliefs based upon unreliable evidence or false assumptions. Frequently they were statements of rather common practices based upon accepted but unreliable evidence. Common

clichés or superstitions would also be included in this type of statements.

- 4.5 *The citing of unreliable authorities.* Statements introduced by phrases such as "Science says . . .," or "People say . . .," or "It is reported in popular magazines that . . ." were used. Here a distinction must be made between such very general or unreliable sources and those which might be used with considerable assurance. However, in any case the mere citation of authority did not in any sense explain why a particular point of view was correct; one would need in addition to give the evidence used by this authority to establish his position on the outcome of the problem.
- 4.6 *Ridicule.* This rather common device of students in their explanations suggested that any position contrary to their own could only be held by someone who did not know the facts.
- 4.7 *Assuming the conclusion.* These statements assumed what was to be proved. This was most frequently represented in these tests by essentially repeating the conclusion by rewording it without changing its meaning.
- 4.8 *Teleology.* These statements assume that plants, animals, or inanimate objects are rational or purposive.

An example of the wording of the directions for one of the tests and a sample problem taken from the test follow.

Form 1.3

APPLICATION OF PRINCIPLES

Directions: In each of the following exercises a problem is given. Below each problem are two lists of statements. The first list con-

tains statements which can be used to answer the problem. Place a check mark (\checkmark) in the parentheses after the statement or statements which *answer the problem*. The second list contains statements which can be used to explain the right answers. Place a check mark (\checkmark) in the parentheses after the statement or statements which *give the reasons for the right answers*. Some of the other statements are true but do not explain the right answers; do not check these. In doing these exercises then, you are to place a check mark (\checkmark) in the parentheses after the statements which *answer the problem* and which *give the reasons for the RIGHT answers*.

In warm weather people who do not have refrigerators sometimes wrap a bottle of milk in a wet towel and place it where there is a good circulation of air. *Would a bottle of milk so treated stay sweet as long as a similar bottle of milk without a wet towel?*

A bottle wrapped with the wet towel would stay sweet

- a. longer than without the wet towel. . () a.
- b. not as long as without the wet towel. () b.
- c. the same length of time—the wet towel would make no difference. . . . () c.

Check the statements below which give the reason or reasons for your explanation above.

- | | |
|-----------------|---|
| Superstition | d. Thunderstorms hasten the souring of milk () d. |
| Right Principle | e. The souring of milk is the result of the growth and life processes of bacteria () e. |
| Wrong | f. Wrapping the bottle prevents bacteria from getting into the milk. . . . () f. |
| Wrong | g. A wet towel could not interfere with the growth of bacteria in the milk. . () g. |
| Wrong | h. Wrapping keeps out the air and hinders bacterial growth. () h. |
| Right Principle | i. Evaporation is accompanied by an absorption of heat. () i. |

- | | |
|-------------------------|---|
| Authority | j. Milkmen often advise housewives to wrap bottles in wet towels. () j. |
| Unacceptable
Analogy | k. Just as many foods are wrapped in cellophane to keep in moisture, so is milk kept sweet by wrapping a wet towel around the bottle to keep the moisture in () k. |
| Right Principle | l. Bacteria do not grow so rapidly when temperatures are kept low. . . . () l. |

In formulating statements for these earlier test forms, no consistent pattern was followed. A study of the results obtained by giving Form 1.3 to many science students suggested the desirability of using in each of the testing situations a pattern of reasons which would remain constant throughout the test. It was believed that this would tend to give a greater reliability to the subscores used in interpretation and thus make the interpretations more meaningful. The pattern of reasons to be included was determined through discussions with teachers who had used Form 1.3. They were asked to indicate the types of items in the test which seemed to be most useful in diagnosing students' difficulties. Using their suggestions, tests employing a pattern of responses were constructed by following through these steps: Situations were selected using the criteria described for Form 1.3 but with greater emphasis upon problems of social significance. These situations were worded in a way that would require an explanation, prediction, choice of course of action, or an evaluation of any one of these. Three conclusions were then formulated, one being defensible through the use of science principles as preferable to the other two. In every case the other two conclusions would not be nonsensical, absurd, or preposterous.

The reasons used in the test were arrived at by first supporting the correct conclusion by formulating three statements of facts or principles which support it and by implication eliminate the other two conclusions. Four wrong reasons

which, if accepted as true, would support the other conclusions were next formulated. Two of these would tend to support one of the wrong conclusions and two the other. They would all tend by implication to eliminate the right conclusion. One statement was formulated so as to be true but irrelevant to the explanation of the problem. One each of the following kinds of reasons completed the pattern—a teleological statement, ridicule statement, assuming the conclusion, unacceptable analogy, unacceptable authority, and unacceptable common practice. Each of these was worded to appear to be consistent with the conclusion keyed as right. Tests following this general procedure were constructed for the areas of chemistry (Form 1.31), physics (Form 1.32), biology (Form 1.33), and general science (Form 1.3a).¹⁰

A sample problem taken from Form 1.3a is given with the directions and key.

PROBLEM

The water supply for a certain big city is obtained from a large lake, and sewage is disposed of in a river flowing from the lake. This river at one time flowed into the lake, but during the glacial period its direction of flow was reversed. Occasionally, during heavy rains in the spring, water from the river backs up into the lake. What should be done to safeguard effectively and economically the health of the people living in this city?

Directions: Choose the conclusion which you believe is most consistent with the facts given above and most reasonable in the light of whatever knowledge you may have, and mark the appropriate space on the Answer Sheet under Problem —.

Conclusions:

- ✓ A. During the spring season the amount of chemicals used in purifying the water should be increased. (Supported by 3, 7, 10, 12)
- B. A permanent system of treating the sewage before it is

¹⁰ A junior high school test, Form 1.3j, which uses a somewhat different and less complex technique was also constructed.

dumped into the river should be provided. (Consistent with 5, 8, 12)

- C. During the spring season water should be taken from the lake at a point some distance from the origin of the river. (Consistent with 12, 14)

Directions: Choose the reasons you would use to explain or support your conclusion and fill in the appropriate spaces on your Answer Sheet. Be sure that your marks are in one column only—the same column in which you marked the conclusion.

Reasons:

- | | |
|------------------------------|---|
| False analogy | 1. In the light of the fact that bacteria cannot survive in salted meat, we may say that they cannot survive in chlorinated water. |
| Irrelevant | 2. Many bacteria in sewage are not harmful to man. |
| Right Principle | 3. Chlorination of water is one of the least expensive methods of eliminating harmful bacteria from a water supply. |
| Ridicule | 4. An enlightened individual would know that the best way to kill bacteria is to use chlorine. |
| Wrong Supporting B Authority | 5. A sewage treatment system is cheaper than the use of chlorine. |
| Right | 6. Bacteriologists say that bacteria can be best controlled with chlorine. |
| Wrong Supporting B | 7. As the number of micro-organisms increases in a given amount of water, the quantity of chlorine necessary to kill the organisms must be increased. |
| Assuming Conclusion | 8. A sewage treatment system is the only means known by which water can be made absolutely safe. |
| Right | 9. By increasing the amount of chlorine in the water supply, the health of the people in this city will be protected. |
| | 10. Harmful bacteria in water are killed when a small amount of chlorine is placed in the water. |

- | | |
|------------------------------|--|
| Tele-
ology | 11. When bacteria come in contact with chlorine they move out of the chlorinated area in order to survive. |
| Right
Supporting
A B C | 12. Untreated sewage contains vast numbers of bacteria, many of which may cause disease in man. |
| Prac-
tice | 13. In most cities it is customary to use chlorine to control harmful bacteria in the water supply. |
| Wrong
Supporting C | 14. Sewage deposited in a lake tends to remain in an area close to the point of entry. |

An examination of the complete test would show that the problem situations included in this form of the test deal with personal health, public health, eugenics, conservation, and the like, and many of them involve questions of opinion as well as of the operation of science principles. The desirability of using these types of problem situations was mentioned by many of the science teachers who had used the earlier form of the test; however, after such problems were formulated it was discovered that very little agreement could be secured among these teachers as to the most defensible conclusions for such problems. This difficulty is illustrated by the above problem on water supply. Several science principles might be cited in proposing a solution to the problem of securing for this city a supply of water free from pathogenic bacteria; but whether or not a supply of water free from pathogenic bacteria constitutes an "effective" safeguard of the health of these people and whether or not any proposed method of securing such a supply of water will be "economical" cannot be determined by science principles alone.

In choosing any one of the three conclusions given with this problem, it is necessary for the student to interpret the terms *effectively* and *economically*. If the student regards reasonable safety, such as might be secured by the adminis-

tration of additional chemicals to the water supply, as an effective safeguard, and if he regards the use of chemicals as an economical practice, then he might defend conclusion A. However, another student might wish to defend conclusion B by pointing out that the use of chemicals assures only a reasonable safety under ordinary conditions and may fail under unusual circumstances, such as the sudden reversal of flow of the river, and that this practice cannot be considered economical in the long run when all the benefits of a sewage disposal system are considered. Still another student might defend conclusion C as representing a more effective safeguard than that of A and a more economical practice than that of B.

The difficulty of keying any of these responses by students as the correct one, unless one knows all of the evidence and values which the student would use to support his point of view, is obvious. Insofar as the student considers the probable effects of these practices upon the people living in the city, upon the people in nearby regions or in towns lying along the river, upon the future as well as the present citizens of this region, and upon the biological life in the waters of this region, he may interpret the terms *effectively* and *economically* so as to justify any of these three conclusions. The pertinent science principles can only aid a person in predicting the effects of each of these practices; they cannot determine whether or not these effects are to be desired. Other students might wish to remain uncertain about which conclusion to choose until further evidence had been obtained about the problem. Such evidence might reveal that it would be better to put into practice all three of the suggested conclusions, i.e., purify the sewage by a permanent system of treatment before it is dumped into the river, take the water from the lake at a greater distance from the shore, and finally add chlorine to the water before it is put into the water mains. It should be clear from this discussion that

the effort to construct a test form which involved social values as well as scientific principles led to situations which were well suited for generating a desirable type of thinking, but which at the same time created considerable technical difficulty for the test constructors. In the discussion of the next test in this series a method for solving these difficulties, at least partially, will be discussed.

Structure of Form 1.3b

In developing Form 1.3b two changes were made: (1) the adoption of a different form of conclusion and the consequent inclusion of reasons to be used if the student were uncertain about the conclusion; (2) addition of acceptable analogy and acceptable authority to the reasons to be used to support or refute the conclusion. A keyed sample problem from Form 1.3b is reprinted here to illustrate these changes:

PROBLEM I

A motorist driving a new car at night at the rate of 30 miles per hour saw a warning sign beside the road indicating a "through highway" intersection 200 feet ahead. He applied his brakes when he was opposite the sign and brought his car to a stop 65 feet beyond the sign. Suppose this motorist had been traveling at the rate of 60 miles per hour and had applied his brakes exactly as he did before. He would have been unable to stop his car before reaching the "through highway" intersection.

Directions:

- A. If you are uncertain about the truth or falsity of the underlined statement, place a mark in the box on the answer sheet under A.
- B. If you think that the underlined statement is quite likely to be true, place a mark in the box on the answer sheet under B.
- C. If you disagree with the underlined statement, place a mark in the box on the answer sheet under C.

Directions for Reasons:

If you placed a mark under A, select from the first ten reasons given below all those which help you to explain thoroughly why you were uncertain and place a mark in Column A opposite each of the reasons you decide to use.

If you placed a mark under B, select from reasons 11 through 24 all those which help you to explain thoroughly why you agreed with the underlined statement and place a mark in Column B opposite each of the reasons you decide to use.

If you placed a mark under C, select from reasons 11 through 24 all those which help you to explain thoroughly why you disagreed with the underlined statement and place a mark in Column C opposite each of the reasons you decide to use.

Reasons to be used if you are uncertain:

- | | |
|----------------------|--|
| Lack of Experience | 1. I have never driven an automobile at 60 miles per hour and don't know how far an automobile will travel after the brakes are applied. |
| Irrelevant "Control" | 2. The distance required to bring a car to a stop depends upon the condition of the road surface. |
| Irrelevant "Control" | 3. The reaction time of the driver is an important factor in determining the distance a car will travel before it stops. |
| Irrelevant "Control" | 4. The mechanical efficiency of the brakes will affect the distances required for stopping a car. |
| Irrelevant "Control" | 5. Whether the brakes are of the mechanical or hydraulic type would make a difference in the stopping distance. |
| Irrelevant "Control" | 6. There are too many variable conditions in the situation to enable one to be sure about the stopping distance. |
| Lack of Knowledge | 7. I do not know which mathematical formula to apply in this problem. |
| Irrelevant "Control" | 8. The distance required to bring a car to a stop depends upon the mass of the car as well as the speed. |
| Irrelevant "Control" | 9. Whether he stopped the car or not before entering the intersection would depend upon how good a driver he was. |

- Irrelevant "Control" 10. The condition of the tires would be a factor to consider in determining the stopping distance for the automobile.

The description of this problem includes an underlined conclusion which the student is asked to judge. The student may agree, disagree, or be uncertain about the conclusion. In the earlier tests he had been asked to select from a list of conclusions the one he thought most appropriately answered the question asked in the description of the science situation. The use of this form of the problem was adopted in order to score the student on his ability to distinguish between problems in which sufficient information was given to enable him to be reasonably sure of his answer, and others about which he should remain uncertain because necessary information was not included in the description of the problem. This form of the problem also enables the teacher to discover those students who have become "over-critical," i.e., who challenge problems by choosing the uncertain response when, in the judgment of the teachers, these problems are so stated that one can either agree or disagree with the conclusion.

An investigation was undertaken to discover what effect the changed form of presenting the conclusion might have upon the results. It was found that it made little difference in which form the conclusion was given. Ninety-one students were given a test especially prepared for this investigation in which they were asked to select from a list of four conclusions the one that they believed was most appropriate. This was followed in the same testing period by a second prepared test in which they were asked to make a judgment about a single conclusion. Two sample items are given here to illustrate how the problems were paired in the two tests.

TEST I, PROBLEM I

A motorist had his tires filled to 35 pounds of pressure when the temperature was 110° F. The temperature dropped to 80° the

next day. What probably happened to the pressure of the air in the tires? (Assume that no air is lost from the tires.)

- () A. The pressure would be greater than 35 pounds.
- () B. The pressure would be less than 35 pounds.
- () C. The pressure would not change.
- () D. The pressure may be the same, greater, or less—one cannot tell.

TEST II, PROBLEM I

A motorist on a trip to the West had his tires checked to 35 pounds on the edge of Death Valley Desert at about 4:00 P.M. That night he stayed at a nearby tourists' camp where the temperature always dropped several degrees during the night. In order to be sure that the old tires on his car would not blow out during the night, *he should let some of the air out of the tires.*

- () Agree () Disagree () Uncertain

Twenty-two such paired problems were included in the two tests. A correlation between the number of right responses made on the two tests was found to be .83. The two tests were found to be about equally reliable (.53 and .55). The mean of test I was slightly higher (10.91) than the mean of test II (10.02) indicating that it was slightly less difficult. The responses of the individual students to the paired problems on the two tests were found to be consistent in 75 per cent of the cases. From this study it seems likely that a score obtained from a test in which the student is asked to select a conclusion for a stated problem will be a good index of his score on a test in which he is asked to judge a given conclusion. Because the student is required to do less reading and consequently can react to more problems in a given unit of time, the type of problem requiring a judgment about a single conclusion was adopted.

The introduction of the "uncertain" response required a new list of reasons to be included (reasons 1 to 10). These ten reasons enable the student who chooses the uncertain

response to explain why he is unable to agree or disagree with the conclusion. Most of these reasons are statements of additional factors which one might want to know before making a decisive judgment about the conclusion. They have been called "control" statements in the problems where uncertainty is considered the acceptable response to the conclusion, and "irrelevant controls" in those problems where either agreement or disagreement with the underlined conclusion is considered the acceptable response. It was also recognized that one might be unable to agree or disagree with the conclusion because of insufficient knowledge about the problem. To provide for this, statements which enable the student to say that he is unable to make a decision because of lack of knowledge about, or experience with, this sort of situation are included in the first ten reasons.

The student who chooses the uncertain response to the problem marks only those of the first ten reasons which he selects to explain his uncertainty and then proceeds to the next problem. The student who agrees or disagrees with the underlined conclusion disregards the first ten reasons and selects his supporting statements from reasons 11 to 24. The pattern of reasons included for supporting or refuting the conclusion is similar to that described for Test 1.3a, with two exceptions. These are the inclusion of an "acceptable" analogy and an "acceptable" authority statement in each problem.

Continuation of PROBLEM I (p. 95)

Reasons to be used if you agree or disagree:

- | | |
|--------------------|---|
| Tele-
ology | 11. The increasing difficulty of stopping objects at higher speeds is a part of nature's plan to keep people from driving too fast. |
| Wrong
Principle | 12. The distance required to bring a car to a stop is directly proportional to the speed of the car. (Inconsistent with B) |

- | | |
|------------------------|--|
| Acceptable Practice | 13. Many drivers have learned from experience that the distance required to bring a car to a stop is more than doubled when the speed is doubled. (Inconsistent with C) |
| Unacceptable Analogy | 14. Just as the centrifugal force acting on a car going around a curve is increased four times when the speed is doubled, so will the distance required to stop a car be increased four times when the speed is doubled. (Inconsistent with C) |
| Right Principle | 15. When brakes are applied with constant pressure there is constant deceleration of the car. |
| Ridicule | 16. Any student of physics ought to know that the distance required to stop a car when it is traveling at 60 miles per hour is more than 200 feet. (Inconsistent with C) |
| Assuming Conclusion | 17. It would require more than 200 feet for the motorist to bring his car to a stop traveling 60 m.p.h. (Inconsistent with C) |
| Wrong Principle | 18. As the speed of a car increases, the mechanical efficiency of the brakes decreases considerably. (Inconsistent with B) |
| Right Principle | 19. When the speed of a car is doubled, the distance required to bring it to rest is increased four times. (Inconsistent with C) |
| Unacceptable Authority | 20. Automobile mechanics report that cars traveling at 60 miles per hour cannot be brought to a stop within 200 feet. (Inconsistent with C) |
| Right Principle | 21. The distance moved while coming to rest by an object undergoing constant deceleration is proportional to the square of the velocity. (Inconsistent with C) |
| Wrong Principle | 22. When the velocity of a car is doubled, the distance required to bring it to a stop may be quickly calculated by multiplying the velocity by four. (Inconsistent with C) |
| Right Principle | 23. The kinetic energy of a car traveling at 60 miles per hour is four times that of the same |

- car traveling 30 miles an hour. (Inconsistent with C)
- Acceptable Analogy 24. Just as the penetrating distance of a bullet is increased four times when its velocity is doubled, so is the stopping distance of an automobile increased four times when its speed is doubled. (Inconsistent with C)

In the earlier forms of the test all analogy statements were formulated as unacceptable reasons. In this form two analogy statements are used in each problem, one acceptable as a reason for supporting the conclusion, the other unacceptable. The inclusion of acceptable analogy statements makes it possible to score a student on his ability to distinguish between those statements of situations which are closely analogous to the original problem and those which seem to be but actually are not explainable by means of the same underlying principles. The use of authority and practice had also been restricted in earlier test forms to the unacceptable use of such reasons. Because in life students are often forced through exigencies of time and circumstance to use authority, it was thought desirable to include in this test two such statements in each problem, one of which was judged to be acceptable and the other unacceptable. If students then used such statements in justifying their reaction to the conclusion, one would be able to distinguish those students who used authorities discriminately from those who either did not cite authorities or who were unable to distinguish between authorities judged acceptable and those judged unacceptable. The inclusion of these statements gives students an opportunity to reveal whether or not they can distinguish between authorities—either persons or institutions—which, because of training, study, experience, etc., should be in a position to give reliable information about the problem, and those which involve the use of false credentials, or transfer of prestige from one field to another, and in reality offer little reliable evidence about the problem.

SAMPLE DATA SHEET

School Experimental
Grade 11

Summary for
Test 1.3b

Column Numbers	Conclusions			Uncertain Knowledge Experience	Reasons			Principles			Controls Analogies			Authority Practice			Ridicule, Assuming Conclusion, Teleology			Inconsistency
	Acceptable	Too Uncertain			No.	R	%R	No.	R	%R	No.	R	No.	R	No.	R	No.	%		
		Too	Certain																	
1	2	3	5	7	8	9	11	12	13	15	16	18	19	21	22	24	25	27	28	
Student A	8	0	0	1	31	26	84	15	15	100	5	5	6	4	2	2	3	10	0	0
B	7	0	1	0	61	40	66	24	21	88	6	6	12	7	12	7	8	13	0	0
C	7	0	1	0	34	32	94	24	23	96	5	5	5	5	0	0	1	3	1	3
D	4	2	1	5	61	31	51	20	16	80	13	6	8	4	9	5	12	20	14	23
E	2	4	1	10	20	7	35	5	3	60	6	2	1	0	4	2	4	20	0	0
F	4	2	1	4	27	20	74	12	10	83	5	4	6	4	2	2	3	11	7	26
G	3	2	2	4	70	26	37	28	16	57	12	0	10	5	8	5	13	10	14	20
Maximum Possible	8	6	2	20	106	65	100	56	34	100	60	15	16	8	16	9	24	100	97	100
Low Score	1	0	0	0	4	3	22	0	0	0	0	0	0	0	0	0	0	0	0	0
High Score	8	6	2	11	70	33	100	29	24	100	36	13	13	7	9	7	12	30	21	50
Group Median	4	2	1	4	26	14	56	10	8	86	5	2	3	2	3	2	3	13	1	4

*Summarization and Interpretation of the
Scores on Form 1.3b*

The form of the data sheet on which the several scores are tabulated and summarized is presented on page 102. A description of how these scores are obtained from the test results and some of the possible interpretations is also given below. Some of the experimental procedures used for arriving at this form of summary will also be described.

An experimental form of Form 1.3b was given to 415 students who were in the eleventh and twelfth grades of two large public high schools (161 juniors and 254 seniors). The results were studied in an attempt to discover a convenient and meaningful method for reporting achievement. An item analysis or record of the responses of students to each item in the test was prepared. This was studied to reveal items which seemed to need revision either because they were too difficult, because they were ambiguous, or for some other reason did not elicit the expected student response. A score indicating the number of student responses on each separate kind of item was then put on a tentative data sheet. Twenty-seven scores were used for each student on this original data sheet, and several others were computed from these in an effort to find those which gave the most meaning to the results.

The interrelationships of the scores were also studied. From these preliminary studies the final form was made and given to a new group of 283 students from two schools in the Eight-Year Study. These students included 127 from the tenth grade, 166 from the eleventh grade, and 40 from the twelfth grade. These results were used for the statistical data which will be found in Table 4 of Appendix II.

The final form for reporting scores determined by these means contains 20 scores for each student. These 20 scores seem to give all of the essential information necessary to

describe the differences in the students' ability to apply principles in the manner defined and measured by this test. An examination of the data sheet (p. 102) will show how these scores were finally recorded.

The scores made by seven students in the eleventh grade were selected for purposes of illustration. At the bottom of the sheet the maximum possible score, highest score, lowest score, and group median is recorded for each column. These were computed from the class from which these seven students were selected. Some of the scores represent actual number of responses, while others are computed in per cent by using certain of the scores from other columns as bases.

The achievement of the student as revealed by the test may be analyzed in terms of five related questions. The *first* of these questions is: To what extent can the student reach valid conclusions involving the application of selected principles of science, which he presumably knows, to new situations?

*Columns*¹⁷ *Column 1* gives the number of conclusions out of a possible eight which the student marked correctly. *1, 2, 3* The eight correct responses were distributed among agreement with the stated conclusion in three problems, disagreement with the stated conclusion in three problems, and uncertainty about the stated conclusion in the remaining two problems. *Column 2* (too uncertain) gives the number of conclusions which the student marked uncertain when the correct response was either "agree" or "disagree." *Column 3* (too certain) gives the number of conclusions which the student marked either agree or disagree when the correct response was "uncertain." When his scores in columns 1, 2, and 3 do not total to eight, either the student marked some conclusions agree which should have been marked disagree, or he marked some con-

¹⁷ The column numbers used in the following paragraphs refer to the summary sheet (p. 102) on which the scores are recorded.

clusions disagree which should have been marked agree, or else he omitted some of the conclusions. If we denote an interchange of the agree and disagree responses by the term "error in fact," the following table may be used to describe the complete scoring of the student's conclusions.

Student	Key	Agree	Uncertain	Disagree
Agree		Acceptable	Too certain	Error in fact
Uncertain		Too uncertain	Acceptable	Too uncertain
Disagree		Error in fact	Too certain	Acceptable

Thus on the sample data sheet student A marked all eight of the conclusions in agreement with the key. Student D agreed with the key four times, marked two of the conclusions as uncertain when he should either have agreed or disagreed with them according to the key. He also marked one of the conclusions which was keyed as uncertain as agree or disagree. Further he either made an "error in fact" by marking an agree conclusion as disagree or a disagree conclusion as agree, or he omitted one problem. This is shown by the fact that his score on conclusions totals seven rather than eight. One would have to examine his paper to determine whether he had omitted a problem or made an "error in fact," for no score for problems omitted is recorded on the data sheet.

The *second* question is: How does the student explain his uncertainty when he marks the stated conclusion "uncertain"?

Columns Column 5 gives the number of statements which the student used to express either a lack of knowledge

about, or experience with, the situation described in the problem. These explain why he marked one or more of the stated conclusions "uncertain." These statements are considered neither "right" nor "wrong" in scoring the test. *Column 15* gives the number of statements which express a desire for control (see the test items themselves to clarify the intended meaning of "Control"). They also are used by the student to explain why he marked one or more of the stated conclusions "uncertain." In two of the eight problems there is actually a need for further clarification or control of certain factors involved in the problems. *Column 16* gives the number of statements, used by the student in these two uncertain problems, describing "controls" which are considered to be essential additional information necessary for the solution of the problem, and hence are valid reasons for marking the conclusion uncertain. In the remaining six problems, the controls are considered to be unnecessary for the solution of the problem. The difference between the scores in columns 15 and 16 gives the number of unnecessary controls marked by the student. It should be borne in mind that a student has an opportunity to score in columns 5 and 15 when he marks a conclusion "uncertain," but has an opportunity to score in column 16 only when he marks the conclusion "uncertain" in one of the two problems where the uncertain response is regarded as the correct one.

Student D, as shown in column 5, used five statements which expressed either a lack of knowledge about, or experience with, those problems which he marked as uncertain. Generally speaking, a high score in column 5 will be associated with a low score in column 1. The correlation between column 1 and column 5 is $-.34$. The fact that he has a score of one in column 3 indicates that he marked one of the problems which was keyed as uncertain in agreement

with the key, while the score of six in column 16 indicates that he must have judged correctly the other uncertain problem. His score of two in column 2 would account for the seven unacceptable control statements which were used (difference between columns 15 and 16) for in these two problems he was attempting to justify an uncertainty through the use of "control" statements when according to the key he should have either agreed or disagreed with the conclusion. In summary, student D marked four of the conclusions in agreement with the key. He was too uncertain in two of the problems and too certain in one. He either omitted one problem or made an "error in fact" by marking an agree conclusion disagree or a disagree conclusion as agree. He used five statements to indicate that he did not understand some of the problems where he was uncertain about the conclusion, and thirteen statements of "controls," six of which were considered to be acceptable.

The *third* question is: To what extent can the student justify logically his agreement with, his uncertainty about, or his disagreement with the stated conclusions?

Columns 7, 8, 9, 27, 28 Column 7 gives the total number of reasons used by the student to explain his decisions about the stated conclusions (excepting those which express a lack of knowledge about, or experience with, the situation described in the problem scored in column 5). Students vary a great deal in their comprehensiveness, that is, in the extent to which they use a large number of reasons to explain their decisions about the stated conclusions. The meanings of every subscore on reasons for a chosen student must be interpreted in the light of the score which he received in column 7. Column 8 gives the number of correct or acceptable reasons used by the student. Column 9 gives the per cent accuracy of the student in supporting his decisions about the stated conclusions with acceptable reasons.

puted by dividing the score in column 12 by the score in column 11 and expressing the result in per cent. The scores in column 16 were discussed above in connection with the second question. *Column 19* gives the number of "sound" analogies used by the student. The difference between the scores in columns 19 and 18 gives the number of unacceptable or false analogies selected by the student. *Column 22* gives the number of acceptable appeals to authority or common practice which the student used in explaining his decisions about the stated conclusions. The difference between the scores in columns 22 and 21 gives the number of unacceptable appeals to authority or common practice selected by the student.

Student C used a total of 34 reasons to justify the eight conclusions he selected. Twenty-four of these were restricted to principles, of which 23, or 96 per cent, were keyed as acceptable. He also used five acceptable analogies, and only one statement which was classified as unacceptable because it was a ridicule, teleological, or assuming the conclusion type of reason. He did not use authority or common practice to explain his choice of conclusions.

In making interpretations of a student's scores, all of his scores on reasons should be judged in relation to his score in column 7. Per cent scores should be judged in relation to the number of items on which the per cent is based. That is, one out of two may have quite a different meaning than 10 out of 20. Reference to the "maximum possible," the "lowest score" and "highest score," and the group median (all given at the bottom of the summary sheet) will provide a frame of reference for judging the student with respect to the members of his own class.

Statistical data, including the reliability of each score, the intercorrelations of various scores, means, and standard deviations for several populations will be found in the Appendix II, Tables 4 and 5.

If students have been placed in situations in the classroom and laboratory where resourcefulness, adaptability, and selective thinking have been essential for the solution of problems, and if the emphasis given to teaching science principles has been upon their applications to the solution of problems involving commonly occurring natural phenomena rather than on the mastery of science information as an end in itself, then students should have little difficulty in behaving in the manner anticipated by this test. Such students would have had many opportunities to apply the principles of science as they learned them to a number of situations in the laboratory and classroom, and would have been encouraged to be alert for similar opportunities for application as they occur outside the classroom.

Experience of teachers with this objective seems to indicate that the objective is not attained through any one particular teaching unit. Rather it is the outcome of the way in which emphasis has been given to the objective with all the science materials taught in the classroom and laboratory. Consequently, teachers may wish to use from time to time during the semester or year classroom exercises which can be used for checking on these abilities and giving a tentative appraisal of progress. A considerable number of such exercises, much simpler in form than the tests of Application of Principles, have been constructed by classroom teachers in summer workshops.

III. APPLICATION OF PRINCIPLES OF LOGICAL REASONING^o

ANALYSIS OF THE OBJECTIVE

The phrase "logical reasoning" is currently used to describe a wide variety of behaviors. The whole process of thinking about problems in an orderly scientific fashion is sometimes called logical reasoning. In what follows the phrase "logical thinking" will be restricted to mean distin-

guishing between conclusions which follow logically from given assumptions and conclusions which do not follow logically from the given assumptions.

The intended meaning of the term "principles of logical reasoning" may be illustrated by means of the following examples of such principles:

- A. *Definitions*: Crucial words and phrases must be precisely defined, and a changed definition may produce a changed conclusion although the argument from each definition is logical.
- B. *Indirect Argument*: The validity of an indirect argument depends upon whether all of the possibilities have been considered. If there are three and only three possibilities and one of them must happen, then if two of the possibilities are shown to be in fact impossible, the third must happen. The conditions necessary for the logical use of indirect argument are seldom fulfilled in practice.
- C. *Argumentum ad Hominem*: An attack upon certain aspects of a person or institution, even though justified, is not sufficient to prove the lack of all merit in that person or institution. This covers the common use of ridicule, attack on motives, etc.
- D. *If-Then*: If one accepts certain premises, then one must accept the conclusions which follow from these premises. The if-then principle is a necessary part of our method of criticizing generalizations, questioning assumptions, etc.

The belief that the study of certain secondary school subjects develops a faculty for logical reasoning is no longer considered tenable. It is, however, quite different to claim that properly guided contact with the subject matter of the secondary school curriculum may provide experiences which will promote logical thinking in dealing with life situations. Many secondary school teachers are endeavoring to have their students recognize patterns for logical thinking in the organization of certain bodies of subject matter. Sometimes the teachers make a conscious effort to have their students

apply these patterns for thinking to problems which arise in connection with their daily experiences. It is found that principles of logical thinking may be stated and applied to widely different kinds of situations. In the light of the foregoing explanation, the objective under consideration may be stated in general terms as follows: Students in secondary schools should acquire the ability and the disposition to apply principles of logical reasoning in dealing with their everyday experiences.

Several more specific behaviors which might be chosen to characterize progress toward the achievement of the objective are listed below:

- a. Disposition to examine the logical structure of the arguments and to apply principles of logical reasoning in the study of these arguments.
- b. Ability to distinguish between conclusions which do and ones which do not follow logically from a given set of assumptions.
- c. Ability to isolate the significant elements in the logical structure of an argument as shown by distinguishing between statements of ideas which are relevant and statements of ideas which are irrelevant for explaining why a conclusion follows logically from given assumptions.
- d. Ability to recognize the application of a logical principle, whether stated in general terms or specifically referred to the situation in question, to explain why a conclusion follows logically from given assumptions.

No effort to prepare objective tests to measure the disposition of students to apply logical principles in dealing with their everyday experiences was made by the Evaluation Staff. A test devised for this purpose would present serious problems of validation. The difficulties attendant upon the construction and administration of such a test would probably be greater than the difficulties of observing the students directly. Hence the efforts to measure behaviors re-

lated to the objective have been directed toward measuring the abilities connected with applying logical principles rather than toward measuring the *disposition* to apply logical principles.

The following discussion deals with the evaluation of the ability to judge the logical structure of arguments presented in written form. This ability will have much in common with the ability to judge the logical structure of arguments presented verbally, pictorially, or otherwise. Some students will have occasion in later life to write essays, prepare speeches, and the like. For these students an emphasis upon the producer aspect of applying logical principles is easily justified. Almost all students, however, will read editorials and advertisements, listen to political speeches, and the like. Hence this consumer aspect of applying logical principles (for example, taking note of the need for definition of terms) may be considered an objective of general education.

THE DEVELOPMENT OF EVALUATION INSTRUMENTS

Preliminary Investigations

The first step toward the construction of a test for this objective was the preparation of a list of logical principles which secondary school students might be expected to apply. A few principles were found explicitly stated in secondary school textbooks (particularly of geometry) and the list was extended by reference to books on logic. From this list the four stated above were selected. Teachers of mathematics were particularly concerned with the objective, and their interests largely determined the choice which was made. The principles stated relative to definitions, indirect (or *reductio ad absurdum*) argument, and "if-then" reasoning play an important role in the teaching of geometry. The fallacy of *argumentum ad hominem* was included because the claim has so frequently been made that the study of geometry,

which as usually taught offers little opportunity for this sort of error, provides a standard of comparison for reasoning in other situations. Consequently if the acquaintance with this standard is functional, it should enable the student to recognize the fallacy.

The second step toward the construction of a test consisted of a search of current newspapers, magazines, and legal case-books for suitable reasoning situations. These sources were chosen because of the emphasis being given in several of the schools upon reasoning in life situations. The legal cases which formed the basis of several test problems were typical of those reported almost daily by the press, but were believed to be of greater interest to students.

Construction of Early Short-Answer Forms

The first test which was constructed (Form 5.1) described 12 different reasoning situations or problems¹⁸ each followed by several possible conclusions. The student was asked to select one of the conclusions and to defend it by selecting reasons from a list which followed. Each logical principle could be correctly used to defend a conclusion in three different problems. Included in each list of reasons were statements of several of the principles listed above, and additional statements which were irrelevant or otherwise unsatisfactory as reasons. The occurrence of several of the principles in each list of reasons required the student to discriminate among them even if the relatively abstract form of statement helped him to identify them.

In order to discover what sort of statements other than principles should be included among the reasons, a form was prepared which contained only the situations and the several alternative conclusions. Four classes of tenth and eleventh grade students took this test and wrote out their reasons in essay form. Many of the reasons ultimately used in

¹⁸ For a similar problem taken from a later form, see p. 119.

the short-answer form were taken with practically no changes from student papers. This preliminary investigation also served to suggest revisions in the statement of the situations and the conclusions.

The scoring plan finally adopted for Form 5.1 allowed two points for each correct conclusion, one point for each correct reason, and deducted one point for each incorrect reason. A score was given indicating achievement relative to each principle separately, and also a total score.

The next form (5.11) of the Application of Principles of Logical Reasoning test incorporated several changes. It was noted that the statements of logical principles in Form 5.1 were of two kinds. Some of the statements referred directly to the situation under consideration and others were general statements of logical principles. A pattern of statements was built into Form 5.11 with a view to securing separate scores on ability to recognize the application of principles which were stated specifically and principles which were stated generally. In each problem there were four specific statements of principles, four general statements of principles, and two extraneous statements including in the test as a whole statements of personal opinion, prejudice, reliance upon authority, and the like. Of the four specific and four general statements in each problem, one of the specific and one of the general statements were relevant in the sense that they explained why the correct conclusion followed logically from the given assumptions. In a sense the cards were stacked against the student by providing three opportunities to use an irrelevant statement of a principle and one opportunity to use an extraneous statement for each opportunity to use a relevant statement of a principle.

The four principles (definition of terms, indirect argument, *argumentum ad hominem*, and if-then) tested in Form 5.1 were again tested in Form 5.11. In addition, a principle relative to sampling ("A sample does not necessarily repre-

sent the population from which it was drawn") was included in Form 5.11. Three problems on each principle were given, or 15 problems in all.

When the test results were summarized, an attempt was made to score the number of correct conclusions (out of a possible three) on each principle and the number of correct (out of a possible six) and incorrect (out of a possible eighteen) uses of statements of each of the given principles. These scores were found to be too unreliable to be useful in practice. Moreover, the attempt to summarize separately the right and wrong uses of specific and of general statements of principles did not yield results of practical significance. It was found that the scores on specific statements were highly correlated with the scores on general statements.

In the final analysis the scoring of Form 5.11 yielded six useful scores. These were scores on numbers of right and wrong conclusions, right and wrong total reasons, extraneous reasons, and general accuracy. The general accuracy score was computed as twice the total number of right responses (conclusions and reasons) minus the total number of wrong responses (conclusions and reasons). This score was highly correlated with each of the other scores, and a reliability coefficient of .94 was obtained for a population of 216 students.

A consideration of the desirable improvements to be made in revising this test led to several suggestions. Form 5.11 was a long test and was made inefficient by the large proportion of wrong statements. The student who responded correctly to the test problems made an explicit response to only one statement in five. The assumption that by refraining from marking a statement a student was making an explicit response (e.g., "the statement is irrelevant") was not thought to be tenable. Thus the student who refrained from marking a statement might have done so because he did not understand the statement or did not take time to consider it fully

Hence a tentative revision of Form 5.11 was made and given to 60 students. In this form, 5.11a, the students were asked to respond to every statement and to decide whether it was (1) specific and relevant, (2) specific and irrelevant, (3) general and relevant, (4) general and irrelevant. This attempt to get at possible differences in the ability of students to deal with specific and general statements of logical principles was again not successful. No very meaningful interpretations of difference between the ability to deal with specific and the ability to deal with general statements could be made. However, when scored in terms of relevance alone, for example, total number of irrelevant statements classified under (2) or (4) above, Form 5.11a yielded very promising results. With only eight problems based on four principles, it was possible to secure a number of diagnostic scores including scores on each of the principles separately. For this latter purpose the method formerly used for scoring the separate principles on Form 5.11 was changed. Rather than counting the number of correct and incorrect uses of each principle throughout the test, the plan was now adopted of scoring two intact problems both directed at the definition principle to secure a score on accuracy with definition, and similarly with the other principles. This plan was later used in summarizing the results on the final test, Form 5.12. The scoring of this test will be discussed in some detail in what follows.

*Structure of the Application of
Principles of Logical Reasoning Test, Form 5.12*

It has been found that Form 5.12 of the Logical Reasoning test provides a better analysis of the students' abilities in relation to the objective than did previous forms. Moreover, with the exception of the original Form 5.1, this form is considerably shorter than previous forms and somewhat simpler from the standpoint of the directions to the student.

A study of the following explanation of the structure of the test problems in comparison with the sample test problem presented below will serve to clarify the objective further and to indicate the extent to which it is measured by the test. A list of the responses accepted as correct by a jury of competent persons (i.e., a test key) is given in the margin.

Problem IV

In January, 1940, Commissioner K. M. Landis submitted a plan to give financial aid to minor league baseball teams to restore fair competition by preventing certain major league teams from controlling the supply of players. Several leaders in the baseball world objected to this plan; some declared that Landis should enforce the rules governing the operation of baseball teams, but should not make interpretations which would change the intended meaning of the rules set up by the proper committees.

Larry MacPhail, president of the Brooklyn Dodgers, speaking at a dinner in Boston, expressed grave concern over the situation. The following statements are quoted from his remarks: "In the matter of Landis versus the present system, he sits as prosecutor, judge, and jury, and there is no appeal. If baseball is to be dominated by any selfish group, it won't be long before professional football or some other sport will replace baseball as the great national game, and none of us want that."

Directions: Examine the conclusions given below. If by "us" Mr. MacPhail means all persons at the dinner, and if they accept his remarks as true, which one of the conclusions do you think is justified?

Conclusions

- A. Logical persons at the dinner will conclude that they do not want baseball to be dominated by a selfish group.
- B. Logical persons at the dinner will conclude that, if the domination of baseball by a selfish group is prevented, baseball will not be replaced as the great national game.
- C. It is impossible to say what a logical person at the dinner will conclude.

A: *Statements which explain why your conclusion is logical.*

Mark in column B: *Statements which do not explain why your conclusion is logical.*

C: *Statements about which you are unable to decide.*

Statements

- A 1. Since we assumed that Mr. MacPhail referred to all persons present at the dinner when he said "none of us," and that those present accepted his statements as true, the conclusion which we reached follows logically.
- B 2. Logical persons at the dinner may agree or disagree with Mr. MacPhail.
- B 3. Without knowing the assumptions of logical persons, we cannot predict their conclusions.
- A 4. If no person at the dinner wants professional football or some other game to replace baseball as the great national game, then the logical ones cannot want baseball to be dominated by a selfish group.
- A 5. If we accept the assumptions on which an argument is based, then, to be logical, we must accept the conclusions which follow from them.
- B 6. Sometimes the meaning of a word or phrase used in an argument must be carefully defined before any logical conclusion can be reached.
- B 7. A changed definition may lead to a changed conclusion even though the argument from each definition is logical.
- B 8. If the domination of baseball by a selfish group results in some other sport replacing baseball, then, if such selfish domination is prevented, baseball will not be replaced.
- B 9. Mr. MacPhail considered every possibility—either baseball will or will not be replaced as the great national game—and thus made a sound indirect argument.
- A 10. If a conclusion follows logically from certain assumptions, then one must accept the conclusion or reject the assumptions.

- B 11. If one removes the fundamental cause for other games replacing baseball, baseball will not be replaced as the great national game.
- B 12. The soundness of an indirect argument depends upon whether all of the possibilities have been considered.

In each problem the student is given a paragraph, three conclusions, and twelve statements. He is directed to read the paragraph carefully and to choose the one of the three conclusions which he thinks is justified by the paragraph. In the test as a whole the student judges the logical appropriateness of the conclusions drawn in eight different situations. In two of these the definition principle operates; in two others the indirect argument principle operates; in two others the *argumentum ad hominem* principle operates; and in the remaining two the if-then principle operates. It should be noted that the number of possible correct conclusions is small, especially if considered with respect to the opportunity to use the correct principles separately. Consequently the major emphasis is placed upon the students' reactions to the statements which follow the conclusions in each test problem.

The statements offered to the students are of several kinds, including:

- a. General or abstract statements of the logical principle involved in that particular test situation.
- b. Specific statements of the logical principle involved in the particular test situation.
- c. General or specific statements of logical principles not pertinent to the particular test situation, statements which appeal to authority, statements of personal opinion, or statements which are otherwise irrelevant.

The student is directed to mark each statement in one of three ways according as it is:

- a. Relevant for explaining why his conclusion is logical.
- b. Irrelevant for explaining why his conclusion is logical.
- c. Not sufficiently meaningful to him to permit a decision.

In the test as a whole the student judges the relevance of 96 statements, and is given the opportunity to reveal his lack of understanding of any of these statements. The variety of the statements including specific and general statements of the principles, statements of authority, personal opinion, prejudice and the like provides an opportunity to make many of the common logical errors. The sample of statements in the test includes 36 relevant and 60 irrelevant statements. Of the 36 relevant statements, 16 are specific and 20 are general. Of the 60 irrelevant statements, 20 are general statements of the four principles of the test, 19 are specific statements of these principles, and 21 are specific statements of the other kinds mentioned above.

Summarization and Interpretation of the Scores on Form 5.12

During the experimental stages of Form 5.12, the test results for a sample population of 351 students were studied intensively in an attempt to discover the most convenient and most meaningful form for reporting the results. An item analysis or record of the responses of all students to each item on the test was prepared. The individual student papers were scored by entering the number of responses of each separate kind on a tentative data sheet. Fourteen scores were summarized for each student, and more than eight additional scores were considered during the study. Certain important scores were selected and studied with reference to the item analysis in an effort to see more clearly the relationships between each of these scores and the responses of students to individual test items.

The 351 students comprised 12 separate classes in four

public schools. Certain facts about the backgrounds of these different classes were known. The responses of each class to the individual test items (taken from the item analysis), and the median scores of each class (taken from the data sheets), were studied in an attempt to discover the degree of agreement or disagreement of these results with the known facts about the various classes of students. The results of this study indicated that the students who secured good total scores were also the students who did well with the individual test items. Moreover, it was found that the classes which had had most contact in school with the logical reasoning objective tended to secure the highest scores on the test.

Certain correlation coefficients between the scores which had been summarized were computed. It was found possible to reduce the number of scores on the data sheet to 11 without an appreciable loss of information. It was again found that separating the responses to specific statements of principles from the responses to general statements of principles did not yield results of practical significance. Several attempts were made to secure a general accuracy score which would serve as a good over-all index of behaviors involved in the application of principles of logical reasoning. For example, the total number of correct responses to statements on the test, and twice the number of relevant statements recognized as such, less the number of irrelevant statements judged to be relevant, were tried. It was found that all of these indices were highly correlated with one or none of the simpler scores obtained by counting the numbers of responses of a certain kind, and that the indices were no more reliable than the simpler scores. Hence no score in general accuracy was retained. Because the number of irrelevant statements on the test is larger than the number of relevant statements (60 as compared with 36), the score on irrelevant statements recognized as such is more reliable than the score on relevant statements recognized as such (.88 as com-

pared with .72). The correlation studies indicated that if a single index for the abilities measured by this test is desired, the score on the number of irrelevant statements judged to be irrelevant is perhaps the best such index among the 11 scores summarized on the data sheet which was finally adopted.¹⁰

Scores on this test may be interpreted in terms of the answers to the following three questions:

1. To what extent can the pupil reach logical conclusions in situations which may involve his attitudes and prejudices?
2. To what extent can the pupil justify his conclusions in terms of certain principles of logical reasoning?
3. How well can the pupil apply each of the four principles of logical reasoning?

By study of the various scores reported on the data sheet, the teacher may obtain evidence relative to each of these questions. Different patterns of behavior analogous to those described for the test on Interpretation of Data are identifiable in terms of the relation of the separate scores to the group averages.

VALIDITY AND RELIABILITY OF FORM 5.12

The construction of Form 5.12 of the Logical Reasoning test was undertaken in the light of two kinds of previous experience. The previous forms of the test had been given to selected groups of students and the test results carefully studied. The criticisms of certain teachers who were endeavoring to promote the logical reasoning objective were available. Sometimes these teachers based their criticisms upon

¹⁰ This data sheet is similar to those presented above for the tests on Interpretation of Data and Application of Principles of Science. For a sample copy and detailed description of the interpretation of scores from this test the reader is referred to the manual, obtainable from the Progressive Education Association.

their experiences in administering the tests and interpreting the test results. Sometimes these teachers had met in groups for the purpose of studying and criticizing the tests. Both the studies of test results and the suggestions made by teachers as individuals or as discussion groups helped the test makers with the construction of test Form 5.12. In particular, the problem situations were chosen with regard for the interests of secondary school students. Most of the problem situations in this test form are taken directly from statements found in the feature articles and in the editorial pages in newspapers. These quotations were edited to some extent to avoid the introduction of extraneous factors such as unnecessary vocabulary difficulties, lack of clear antecedents for pronouns, and the like. The statements regarding the logical structure of the paragraphs which set forth the problem situations were carefully chosen in an effort to make them typical of the kinds of statements which students commonly make when they are discussing the logical structure of such paragraphs. Several readers went over each test problem carefully in an attempt to discover loopholes in its logical structure. Although it is probably quite impossible to construct a lifelike argument to illustrate just one principle of logical reasoning, and express this argument without ambiguity in words, an effort was made to approach this ideal in the test situations included in Form 5.12 of the logical reasoning test.

The studies upon which the scoring of Form 5.12 of the Logical Reasoning test was based were described above. It is important to note that even a carefully constructed test, which actually provides opportunities for the behaviors in terms of which the objective is defined, may become invalid if the system of scoring adopted does not yield scores which present a true picture of the significant behaviors called forth by the test. Hence it should be noted that careful attention was given to the mode of scoring of Form 5.12 of the logical

reasoning test. When conditions of administration are appropriate and when the results are interpreted by a person who is familiar with the objective and the structure of the test, Form 5.12 provides a measure of a range of significant behaviors related to the logical reasoning objective.

For the purpose of statistical analysis, the scores of 351 students, of whom 292 were finishing grade ten, 28 were in grade eleven, and 31 in grade twelve, were used. These students were all attending public high schools when tested and composed nine classes in grade ten, one class in grade eleven, and one class in grade twelve. The statistical data presented in Appendix II on reliability, intercorrelations of scores, and so forth, Table 6, are based upon a study of these 351 students. Within certain definite limitations these data would apply to other groups of students in the tenth, eleventh, and twelfth grades.

The statistical constants presented will provide enough basic information to enable the teacher trained in statistics to study the significance of changes in the mean scores of a class or in the scores of an individual student.

Form 5.12 of the test on the Application of Certain Principles of Logical Reasoning is recommended only for classes where conscious attention has been directed toward logical reasoning. Otherwise, the students are apt to wonder why they should attempt to reach logical conclusions which are sometimes contrary to their "better judgments." The judgment of the teacher as to the readiness of his class for problems of the type included in the test is for this reason very important.

IV. THE NATURE OF PROOF

ANALYSIS OF THE OBJECTIVE

In the past, teachers of several of the subject fields in the secondary school curriculum have been concerned with

particular aspects of "proof." For example, one of the objectives for courses in demonstrative geometry is to develop an understanding of the meaning of proof, and students in such courses have been expected to learn to prove theorems of geometry. Teachers of courses in which oral and written expression is emphasized have also been concerned with certain aspects of proof. Logical organization has been sought in themes and speeches. Courses in science have relied heavily upon laboratory experiments to "prove" certain laws, and students have been expected to learn to cite experimental evidence for their conclusions. Similarly, teachers of other subject-matter fields have objectives related to the concept of proof, in each case with connotations rather specific to their own field. The following paragraphs present a generalized definition of an objective which has come to be called "the nature of proof."

Both children and adults in our society are constantly bombarded with "proofs"; i.e., by arguments designed to convince them that they should act in certain ways or should believe in certain things. The whole field of advertising directs its efforts toward convincing people to act in certain ways. Children of elementary school age are persuaded by a radio announcer to ask mother to buy a certain brand of breakfast food. Newspapers and magazines contain cartoons which set forth the dramatic stories of lives set right by buying and using a different brand of soap. The editorial pages encourage readers to adopt one of several possible courses of action. Even the news articles in the daily papers are likely to reflect the policy and convictions of the management, and hence may be said to be one of the kinds of "proofs" with which people are bombarded. The books and magazines they read, the plays and movies they see, the lectures and radio talks they hear, and the conversations they have with their associates, all play a part in forming the convictions upon which the actions of people are based.

In particular, students in secondary schools react to the proofs which they meet in their daily experiences. Authorities on the secondary school curriculum and classroom teachers have expressed concern with the problem of guiding the reactions of the students to these proofs. This concern has led many teachers to attempt to have students become critical of proofs and to have students acquire the abilities needed for analyzing proofs. It would be ineffective to have students become critical of the proofs which they encounter unless the students also acquired some of the abilities needed in analyzing proofs. On the other hand, the ability to analyze proofs is not likely to function unless there is a disposition to analyze proofs when the need for such analysis arises. Hence the nature of proof objective should include the ability to judge proofs, and also the disposition to apply this ability on appropriate occasions.

It should be noted explicitly that any of the physical senses may be the medium for arriving at proofs. Touch, taste, or smell may be the basis for simple proofs. The question, "Are the potatoes salty?" is easily answered; the method is to taste them. Sometimes visual impressions also provide simple and direct proofs, but often these impressions involve more subtle factors. The hand may be quicker than the eye; the story told by the moving picture may create certain impressions which lead up to an intended conclusion through a series of inferences. Verbal presentations such as speeches and debates are also common vehicles for proof. The written "proofs" which are so frequently met in daily life have much in common with proofs in the other forms. It is with arguments or proofs presented in written form that we shall be chiefly concerned in this chapter.

One of the important characteristics of proofs should be noted immediately. Some proofs proceed mostly from stated opinions or convictions. Other proofs are based in part upon data derived from experiments or investigations. Both of

these kinds of proofs will involve certain basic assumptions which may be more or less tenable. Whatever the subject matter with which a proof deals, and whatever the form of presentation in which the proof appears, *the location and appraisal of the basic assumptions upon which the soundness of the proof depends becomes a fundamental ability connected with analyzing proofs.*

In the light of the preceding remarks, some of the behaviors which might be chosen to characterize progress toward the achievement of the nature of proof objective are listed below:

- a. Disposition to analyze proofs critically.
- b. Ability to recognize the basic assumptions upon which a conclusion depends, and to see the logical relationships between these assumptions and the conclusion.
- c. Recognition of the need for further data to confirm, qualify, or negate the available evidence.
- d. Ability to distinguish between assumptions whose tenability could be checked by collecting further data and assumptions whose tenability could not be checked in this way. Examples of assumptions of the latter sort are value judgments, statements of preference, and definitions of terms.
- e. Recognition of the possible ways for studying a problem further, and ability to distinguish between fruitful and unfruitful methods of further study.
- f. Willingness to accept or reject assumptions tentatively, and to test the conclusions which follow from these assumptions by acting upon them.
- g. Recognition that new evidence upon the soundness of one or more of the assumptions may make it desirable to reconsider the argument and perhaps to qualify the conclusion tentatively reached.

The efforts of the Evaluation Staff to measure behaviors relative to the Nature of Proof objective were directed to-

ward measuring the *abilities* connected with analyzing written arguments rather than toward the *disposition* to analyze arguments critically. Even when the problem was reduced to measuring the skills involved in the critical analysis of arguments, it was found to be an extremely complex problem. Groups of teachers in the Eight-Year Study were enthusiastic in their approval of the objective, and they suggested many behaviors which seemed to them significant. The task of clarification and simplification was much greater than was originally anticipated. The early forms of the test used experimentally in an attempt to secure insight into the nature of proof objective were too complicated for practical purposes. The persons who worked on this problem were, however, convinced that the objective is very significant for general education at the secondary level and that a continuing effort to overcome the obstacles set up by its complexity is worthwhile.

THE DEVELOPMENT OF EVALUATION INSTRUMENTS

The first nature of proof tests which were constructed presented the student with a described situation which presumably led to a conclusion, and he was asked to write down the assumptions which seemed to him to underlie the argument.²⁰ An analysis of the responses indicated that for the most part they could be classified into a few types. For example, a uniqueness assumption is often needed to clinch an argument—an assumption which states that a product advertised, or a chemical used in an experiment, etc., is the *only* one which has a given property.

The student responses and the results of the analysis were utilized in the construction of a short-answer form. A list of statements relative to a problem situation was given, includ-

²⁰ Cf. H. P. Fawcett, *The Nature of Proof*, *Thirteenth Yearbook* of the National Council of Teachers of Mathematics (New York, Bureau of Publications, Teachers College, Columbia University, 1938), Appendix, Part I.

ing some which purported to represent facts and others which were assumptions. Students were asked to distinguish facts from assumptions, to reconstruct the argument by using statements from the list, and to indicate whether they would accept or reject the conclusion of the reconstructed argument.

The results from the first short-answer form threw a good deal of light on the thinking of the students. Difficulties frequently arose, however, with respect to the use of the terms "fact" and "assumption," and the first part of the test did not discriminate well among students. The scoring of the reconstructed arguments also caused difficulty. The test was therefore revised several times, but limitations of space prevent a discussion of the resulting experience. Only the forms which the test had taken toward the end of the Study can be described here.

Form 5.21 of the Nature of Proof test incorporated several major changes. An attempt was made to have the students locate the basic assumptions underlying the argument, but the term assumptions was not used in the directions to the student. In each problem a paragraph which presumably justified a conclusion stated at the close of the paragraph was presented. There followed a list of statements. Some of these statements were relevant, in the sense that they described assumptions underlying the argument, and some of them were irrelevant. The students were asked to pick out the relevant statements and to decide which of these might logically be used to support the stated conclusion. In this way the students were given an opportunity to locate basic assumptions, and to recognize the function of these assumptions in an argument, although the word assumption was not used in the test directions.

One of the problems taken from Form 5.21 of the Nature of Proof test is given below. The directions, in a shortened

form, are presented along with the problem.²¹ A list of responses accepted as correct in scoring the test, i.e., a test key, is given in the margin. It should be noted that the test key adopted by a committee of competent persons before the test was given to students was changed to some extent when the test results for a sample group of students were studied. It became apparent that the "C" step in the directions was interpreted differently by the students than by the committee. There were also apparent differences in the interpretation given to the "C" step by students. It should also be noted that there was no decision as to a "correct" response to the conclusion.

Read the problem and then:

- A. Select the statements which *either support or contradict* the underlined conclusion.
- B. Select the statements marked under A which *support* the underlined conclusion.
- C. Select the statements marked under B which you do *not consider satisfactorily established by whatever general knowledge you may have, but which must be included in the argument* if the conclusion is to be completely justified.

Conclusion. According to what seems *most consistent with your analysis thus far*, decide whether you:

- | | | |
|---|---|---|
| A. Are inclined to accept the conclusion. | B. Are very uncertain about the conclusion. | C. Are inclined to reject the conclusion. |
|---|---|---|

Reasons. Select the statements marked under C which *might cause you to reconsider your decision* about the underlined conclusion *if more information were made available to you*. Mark these under D.

²¹ The use of A, B, C, D in the directions below is clarified by the complete directions, by the form of the special answer sheet on which the student makes his responses, and also by a sample exercise explained in the general directions. In the marginal keys below, these letters refer to the columns in which a statement should be marked on the answer sheet.

PROBLEM IX

In a radio broadcast the following story was told: "The people in a little mining town in Pennsylvania get all their water without purification from a clear, swift-running mountain stream. In a cabin on the bank of the stream about a half a mile above the town a worker was very sick with typhoid fever during the first part of December. During his illness his waste materials were thrown on the snow. About the middle of March the snow melted rapidly and ran into the stream. Approximately two weeks later typhoid fever struck the town. Many of the people became sick and 114 died." The speaker then said that this story showed how the sickness of this man caused widespread illness, and the death of over one hundred people.

Statements:

- | | |
|------------|--|
| ABCD | 1. Typhoid fever organisms can survive for at least three months at temperatures near the freezing point. |
| Irrelevant | 2. Good doctors should be available when an epidemic hits a small town. |
| ABCD | 3. Typhoid fever germs are active after being carried for about half a mile in clear, swift-running water. |
| A | 4. There may have been other sources of contamination by waste materials containing typhoid fever germs along the stream or at some other point in the water supply of the town. |
| AB | 5. The waste materials of a person who has a severe case of typhoid fever contain active typhoid organisms. |
| AB | 6. Typhoid fever is contracted by taking the typhoid organisms into the body by way of the mouth. |
| Irrelevant | 7. Only a few people in this town had developed an immunity to typhoid fever. |
| A | 8. Typhoid organisms are usually killed if subjected to temperature near the freezing point for a period of several months. |

- | | |
|------------|--|
| Irrelevant | 9. Sickness and death usually result in a great economic loss to a small town. |
| ABCD | 10. The only typhoid organisms with which the people in the town came in contact were in the water supply. |
| Irrelevant | 11. Vaccination should be compulsory in communities which have no means of purifying their water supply. |
| ABCD | 12. The worker's waste materials were the only source of contamination along the stream. |
| A | 13. There may have been other sources of typhoid fever germs in the town such as milk or food contaminated by some other person. |
| AB | 14. The symptoms of typhoid fever usually appear about two weeks after contact with typhoid germs. |

Several further comments on the structure of this sample problem might be added to those made above. When the student has chosen the statements which he thinks support the stated conclusion, he is asked to decide which of these are essential assumptions whose truth he would question. On the basis of his analysis of the problem, the student is then asked to indicate the degree of his acceptance of the stated conclusion. Finally the student is asked to decide which of the essential assumptions might, in the light of further evidence, make it necessary to reconsider his decision about the stated conclusion.

The relationship between the activities which the students were directed to perform and the definition of the objective in terms of behavior will be apparent to the reader. Under ideal conditions the activities which the student performs might be expected to yield evidence on the students' ability to recognize the basic assumptions in an argument, the standard of proof which the student demands, the student's recognition of the tentative nature of the conclusions which are based upon arguments, and the role of reexamining the

underlying assumptions in order to qualify the conclusions which one reaches. In practice, the results do not yield valid evidence on achievement relative to all of these behaviors. For example, students vary a great deal in the number of statements which they recognize as supporting the stated conclusions. This makes the number of opportunities to challenge assumptions different for different students. A still more serious consideration is the possibility for variation in the interpretation of the test directions from student to student. Such variation was noted particularly in connection with the directions for challenging the truth of the statements which had been marked as supporting the stated conclusions. Moreover, the fact that the various activities which the students are requested to carry out are interrelated, so that failure to perform one step seriously interferes with performing the next step, presents a difficulty in interpreting the results. In this connection the number and complexity of the related activities which the students were asked to carry through proved discouraging to many students.

In the next section a description of the structure of Form 5.22 of the Nature of Proof test in which the attempt is made to avoid some of these difficulties, is presented.

Structure of the Nature of Proof Test, Form 5.22

The progress toward Form 5.22 has involved an attempt to simplify both the procedures which students are asked to carry out and the directions for carrying out these procedures. At the same time there has been an attempt to retain many of the aspects of thinking commonly associated with problem-solving and scientific method.

A study of the following explanation of the structure of the test problems in comparison with the sample test problem presented below will serve to clarify the reasons for the inclusion of each part of the test. A list of the responses ac-

cepted as correct by a jury of competent persons, i.e., a test key, is given in the margin.

PROBLEM III

A science class was studying methods of caring for the skin. The teacher described the following experiment and stated the conclusion which had been drawn from it. "A large bottle of each of the five leading brands of hand lotion was purchased from a drug store. The lotion in each bottle was thoroughly mixed by shaking the bottle for three minutes. Five exactly similar water glasses, one for each lotion, were set in a row on a table, and a piece of filter paper was placed over the open top of each glass. Each brand of lotion was tested by pouring a half teaspoonful of it on the piece of filter paper. For the first brand of hand lotion, drops appeared in the water glass within thirty seconds. The other four brands all took longer than one minute, and two brands failed to filter through at all." This experiment shows that the first brand of lotion is absorbed by the skin more readily than any of the others.

I. Directions: In this part, you are to do two things:
Select all statements which could logically be used to support the underlined conclusion. Blacken the space under A opposite the number of each such statement.

At the same time, select all statements which might make the underlined conclusion less acceptable. Blacken the space under B opposite the number of each such statement.

In this part of the test, your decision about a statement should not be influenced by whether you believe the idea expressed to be true or false.

Statements for I and II:

- | | |
|------------|---|
| A C | 1. The contents of one large bottle of a certain brand of hand lotion are exactly like the contents of any other large bottle of the same brand of hand lotion. |
| Irrelevant | 2. The liquid which is absorbed most readily by the skin is the most effective in softening the hands. |

- B 3. To be absorbed by the skin a hand lotion need not pass through the skin.
- Irrelevant 4. Hand lotions are of doubtful value.
- A C 5. The faster a liquid drips through filter paper the faster it will be absorbed by the human skin.
- A C 6. The pores of the skin are quite similar to the little holes between the fibers of filter paper.
- A 7. Since each bottle was given a thorough shaking, the results for each lotion were typical of the performance of the lotion in that bottle.
- B 8. The "pores" in filter paper are constructed quite differently from the "pores" in the human skin.
- Irrelevant 9. The experiment was probably intended to make sales for some cosmetics manufacturer.
- B 10. Although drops of a liquid appeared in the water glass, certain ingredients of the first lotion may have been retained by the filter paper.
- Irrelevant 11. The speed with which a lotion drips through filter paper is no indication of its effectiveness in softening the skin.
- B 12. Water will penetrate filter paper but is not absorbed by the skin.
- Irrelevant 13. The obvious way to test the five lotions is to try them on the hands of a large group of people.
- A 14. The amounts of lotion placed on each piece of filter paper were very nearly the same.

II. *Directions: Select from the statements already marked under A (the supporting statements) those which you would challenge because you are not convinced they are true enough to be used in supporting the underlined conclusion. Blacken the space under C opposite the number of each such statement.*

III. *Directions: Conclusions A, B, and C are stated below. Choose the one which seems to you to be most consistent with your analysis of the situation described in the problem. In the*

block at the top of the answer sheet, blacken the space A, B, or C to indicate the conclusion which you choose.

Conclusions:

- ✓A. This experiment does not help in deciding which one of the hand lotions would be most readily absorbed by the skin.
- B. The experiment suggests that the first brand of hand lotion is absorbed by the skin more readily than any of the others, but the experiment would have to be repeated several times.
- C. The experiment shows that the first brand of hand lotion is absorbed by the skin more readily than any of the others.

IV. *Directions: Hand lotions are commonly used to replace the oils in the outer layers of the skin which are lost through excessive exposure, washing, and other causes. Hence it may be less important to study the extent to which a lotion penetrates the layers of the skin than to study its effect upon the surface of the skin. The statements presented below describe some activities which have been suggested to study the effectiveness of a hand lotion in keeping the skin soft in the absence of an adequate supply of natural skin oils.*

Select all statements that describe activities which you think would help in studying this effect of a hand lotion upon the skin. Blacken the space under A opposite the number of each such statement.

In this part of the test, your decision about a statement should not be influenced by whether you believe the activity described could actually be carried out.

Statements for IV and V:

- | | |
|------------|---|
| A B | 15. Secure a description of the structure of the human skin. |
| Irrelevant | 16. Find out the names of the companies which manufacture each of the brands of hand lotion used in the experiment. |

- | | |
|------------|--|
| A | 17. Make a precise laboratory analysis of each of several brands of hand lotion to find out the amounts and properties of its principal ingredients, such as vegetable oils, water, etc. |
| Irrelevant | 18. Repeat the experiment several times with the same five lotions and under exactly the same conditions. |
| A B | 19. Set up an experiment in which ten boys and ten girls apply a hand lotion to one hand and no hand lotion to the other hand once each day for a month and compare the results. |
| Irrelevant | 20. Send out a questionnaire to a large number of users of hand lotion to find out which brand is most popular. |
| A B | 21. Use hand lotions regularly on several parts of the body and compare the results. |
| A | 22. Set up an experiment to compare the natural skin oils to the oils contained in hand lotions. |
| Irrelevant | 23. Compare the absorbing power of filter paper and human skin. |
| A B | 24. Look for published information about some of the good and bad effects of using different brands of hand lotion. |

V. *Directions: Select from the statements already marked under A only things which you think you or your class in high school could actually carry out. Blacken the space under B opposite the number of each such statement.*

In each problem the student is given a paragraph which presumably justifies an underlined conclusion stated at the close of the paragraph. This is followed by 14 statements. Some of these statements are relevant in the sense that they describe assumptions underlying the argument, and some of them are irrelevant. Some of the relevant statements might be used to support the underlined conclusion and the remainder of them might be used to contradict it. In the first part of the test the student is asked to decide which of these statements are relevant and to mark them as either support-

ing or contradicting. In making these judgments, the student is to disregard the degree of truth or falsity which he may ascribe to the statements in the paragraph or to the statements listed below the paragraph. He is to judge the relevance of a given statement solely in terms of the context of the argument and to decide whether each relevant statement supports or contradicts the underlined conclusion.

In the second part of the test the student's attention is directed toward those particular statements which he marked as supporting statements. He is asked to indicate the ones which he would challenge because he is not convinced that they are true enough to be used in supporting the underlined conclusion. Since the relevant statements describe assumptions necessary in order to establish the underlined conclusion, in a sense the student is asked in the first two parts of the test to decide which statements are necessary assumptions in the argument, and of these, to choose the ones about which he is uncertain or is in doubt.

In the third part of the test the student is asked to choose one of three stated conclusions. One of these conclusions expresses an acceptance, another, a qualified acceptance, and the third, a rejection of the underlined conclusion. In each problem the student is asked to choose the conclusion which seems to him to be most consistent with his analysis of the problem. In order to agree with the test key, the student should in two problems choose acceptance, in four problems choose qualified acceptance, and in two problems choose rejections of the underlined conclusions.

Parts I, II, and III of the test can be given and scored independently of the remainder of the test, and for some purposes may be considered sufficient. However, besides being able to test a stated conclusion (as in parts one and two) by an examination of the assumptions underlying the arguments which purport to establish this conclusion, it is also important to be able to recognize fruitful lines of further

investigation and to distinguish between types of activities which are relevant to testing the conclusion and those which are not. It may also be considered important for students to learn to judge the practicability of a proposed line of investigation. Parts IV and V of the test were designed to secure evidence on the ability of students to appraise the relevance and practicability of proposals for the further study of a problem.

In the fourth part of the test a significant problem which involves further study of the issues raised in Parts I, II, and III is stated. The student is asked to select from a list of statements those that describe activities which would help him to solve this problem. In making his judgment, the student is not to be influenced by whether he believes the activity described could be carried out in a practical sense.

In the fifth part of the test the student's attention is directed toward those particular statements which he selected in Part IV. He is asked to indicate the ones which he thinks he or his class in high school could actually carry out.

The scores given to students reflect their success or failure in carrying out the procedures in each part of the test. The interpretation of the results depends upon the interpreter's understanding of the structure of the test problems. The usefulness of the test results is in direct proportion to the extent of the interpreter's concern with the objective and his confidence that significant behaviors involved in the objective are actually sampled in the different parts of the test.

Summarization and Interpretation of the Scores on Form 5.22

During the experimental stages of Form 5.22, the test results for a sample population of 307 students were studied intensively in an attempt to discover the most convenient and most meaningful form for reporting the results. These students comprised 12 separate classes divided among the

tenth, eleventh, and twelfth grades. The procedure described previously in connection with the test on Application of Principles of Logical Reasoning was also used in this case.²² Twenty-two scores were summarized for each student, and several additional scores were computed from these during the study. Certain important scores were selected and studied with reference to the item analysis in an effort to see more clearly the relationships between each of these scores and the responses of students to individual test items. Certain correlations between the various scores which had been summarized were run. It was found possible to reduce the number of scores on the data sheet from 22 to 13 without an appreciable loss of information. Scores on per cent accuracy, computed as number of responses marked in agreement with the test key divided by total number of responses of that kind, were tried and abandoned because they were somewhat unreliable and apt to be misleading. Moreover an examination of the scores on various kinds of errors which were also summarized yielded the desired information in a slightly different form. A score on the per cent of the statements keyed as *supporting* and marked by students as *supporting* which the students also marked as *critical* was tried in an effort to secure an index of the "criticalness" of a student. This score was found to correlate highly with a score on critical statements marked by students as critical statements. Hence a score on critical statements marked as critical was used as an index of the tendency of a student who had marked a statement as supporting to challenge its truth. This score when used as an index is not subject to the criticism that it depends upon the number of supporting statements which the student marked as supporting since the effect of this dependence was considered and found to be insignificant.

²² See pp. 122-123.

The scores on this test can be interpreted in terms of the answers to five questions:

1. To what extent does the student recognize relevant phases of an argument, and distinguish between considerations which support and ones which contradict a stated hypothesis or conclusion?
2. To what extent does the student challenge the assumptions underlying an argument, and distinguish between assumptions which, from the point of view of a committee of adults, should and should not be challenged?
3. How do the conclusions reached by the student compare with those reached by the committee who made the test?
4. To what extent does the student recognize the relevance of proposals for the further study of a problem?
5. To what extent does the student judge the relevant activities as practicable, i.e., distinguish between activities which, from the point of view of a committee of adults, are and are not practicable?

By a study of the various scores reported on the data sheet the teacher may obtain evidence relative to each of these questions. It is particularly true of this test that the number of patterns of behavior revealed by the test scores is almost as great as the number of students who take the test. Each pattern should be considered as a unique situation to be interpreted.

VALIDITY AND RELIABILITY OF TEST FORM 5.22

The construction of Form 5.22 of the nature of proof test was undertaken in the light of a good deal of negative and some positive evidence on the behaviors of secondary school students relative to the nature of proof objective. Certain "don'ts" were clearly indicated by experience with previous

forms of the test. For example, in Form 5.21 the dependence of each step upon preceding steps made the interpretation of the test results difficult. At the same time, a number of "do's" were indicated. For example, the realization that the basic assumptions upon which a conclusion depends may be expressed in the form of statements which either support or tend to contradict the conclusion (as opposed to statements which are irrelevant) made it possible to get at the concept of assumptions in operational terms.

In approaching the construction of Form 5.22 of the nature of proof test, a need was felt for another check upon the direct responses of students. The students in a geometry class of a large public high school not participating in the Eight-Year Study were selected for this purpose. The teacher of this class was known to be working actively to improve the achievement of this objective. For purposes of illustration, one of the four test exercises which were given is reprinted below together with the responses which one student made to the questions.

Exercise II

Read the paragraph and then answer the questions which follow. Speed is not at all important. You should take enough time to organize your ideas and to state them precisely.

In an agriculture class the teacher was discussing the importance of the use of fertilizer. He described the following experiment: "Some wheat seeds were planted in two large pots of earth. The seeds were of the same variety, and the soil used had been thoroughly mixed and then divided into two parts, one for each pot. Fertilizer was added to one and not to the other. The pots were then placed side by side in a greenhouse and both regularly and equally watered. At the end of three months the wheat plants in the fertilized pot weighed twenty-five per cent more than those in the unfertilized pot."

The class came to the following conclusion: "*Farmers who use*

this fertilizer on land on which they raise wheat will get larger yields of grain."

1. Indicate your reaction to the *underlined conclusion* by a check mark (✓) in one of the three spaces provided.

After a consideration of this experiment I feel that the underlined conclusion is:

Probably true ✓, Completely uncertain —, Probably false —.

Explain your answer in some detail, that is, tell why you felt that the underlined conclusion was probably true, completely uncertain, or probably false.

"I felt that the underlined conclusion was probably true, because if the fertilizer had been placed in the pot where the wheat seed grew the faster, then that would prove it.

"Especially if the soil had been mixed thoroughly and the pots watered equally each day."

2. What things does the class have to assume (take for granted) if the underlined conclusion is to be considered true? You may include statements of ideas reported in the above experiment and also statements of ideas not actually mentioned. Make a separate statement for each assumption which you wish to point out, and number these statements 1, 2, 3,
- "1. The wheat seeds were the same.
 2. The soil was thoroughly mixed.
 3. The two plants were regularly and equally watered.
- ✓4. All wheat, even from the same crop, grows the same as the rest."
3. Now go back to the statements which you listed under point 2 above. You may feel that some of these statements should not be assumed or taken for granted. Place a check mark (✓) beside the number of each statement which you feel should not be taken for granted.
4. What things occur to you which, if true, would contradict the underlined conclusion? Make a separate statement for each contradictory idea and number these statements 1, 2, 3,

- "1. Two wheat seeds, even from the same crop, would most likely not grow the same, even under the same conditions."
5. In what ways could the above experiment be improved? Make a separate statement for each suggested improvement and number these statements 1, 2, 3,
- "1. Take more than two pots and then let them grow under the same conditions, because the more seeds you use, the more perfect will be your conclusion.
- "2. Take soil from same general location and mix, putting one with fertilizer and one without. Use water equivalent to general rainfall in location from which soil is taken from, and at approximately the same intervals.
- "3. Run tests over a greater period of time."

Several significant observations were made from this investigation. The rather weak responses which the student quoted made to question 1 (the general direction was "explain your answer in some detail") are typical of this sample of students. In response to question 2 some of the students wrote out basic assumptions which went beyond a mere repetition of the statements made in the paragraph. Other students found even more difficulty at this point than did the student whose responses are presented above. The responses to question 3 are dependent upon the responses to question 2 and as a result were significant only for students whose performance on question 2 was satisfactory. The responses to question 4 seldom yielded new ideas not previously expressed in the answers to questions 1 and 2. An appreciable number of the students introduced several new ideas in their responses to question 5. The student whose responses are presented above is an example. In summary, the results were as follows: (1) The general direction "Explain your answer in some detail," does not elicit detailed, comprehensive answers. (2) There is a considerable difference in the minds of some students between locating as-

sumptions upon which a conclusion depends and suggesting ways for improving the argument upon which a conclusion depends. In the light of the first point, we would expect difficulties if we attempted to compare the written responses of students to the general direction "Explain your answer in some detail" to their responses on an objective test. In the light of the second point, it may be worthwhile to include in an objective test two logically equivalent forms of questions relative to underlying assumptions: (a) pick out the statements of underlying assumptions, (b) pick out the statements of activities relevant to improving the argument. The reader will recall from his study of the simple problem that an attempt was made in constructing Form 5.22 of the Nature of Proof test to include questions of these two kinds.

The construction of Form 5.22 of the Nature of Proof test was undertaken by a committee of five persons with the assistance at certain stages of several other persons. The test situations and test directions were viewed critically in the light of all of the available evidence from previous forms of the test. An analysis of the statements made by various students provided helpful suggestions for the construction of statements to be included in the objective form of the test. The kinds of irrelevant statements which the students made were especially helpful in building irrelevant statements which would be used as relevant by an appreciable number of students. The results of the statistical study which is described below indicate that the directions to the students are unambiguous, and that several distinct behaviors are measured by the test. The evidence available to date strongly indicates that, under certain conditions of administration, Form 5.22 of the Nature of Proof test provides a valid measure of a certain range of behavior relative to the nature of proof objective.

For the purpose of statistical analysis, 307 students—115 finishing grade ten, 96 in grade eleven, and 96 in grade

twelve—were selected. These students were all attending public high schools when tested and composed five classes in grade ten, three classes in grade eleven, and four classes in grade twelve. The five classes in grade ten, and one of the classes in grade twelve, were then completing a course which emphasized the nature of proof objective.²³ In the remaining groups there was an awareness of this objective, but less specific attention to it. The results of the study seem to indicate that at the present there would be little advantage in computing grade norms, since the emphasis given to the objective has more influence on the scores than does the grade placement of the students from the tenth to the twelfth grades.

The statistical data presented in Appendix II, Table 7, are based on this population of 307 students. Within limitations these data would apply to other groups of students in the tenth, eleventh, and twelfth grades. If a chosen group is comparable to the sample group, the statistical constants presented in Appendix II, Table 7, will provide enough basic information to enable the teacher trained in statistics to study the significance of changes in the mean scores of a class or in the scores of an individual student. The reliabilities of the various scores are in general not as high as have been obtained in other tests of thinking abilities. A number of the scores are, however, fairly reliable, and it is a reasonable hypothesis that the interpretations drawn on the basis of a careful examination of the *patterns* of scores are more trustworthy than the reliability of the separate scores would suggest.

A RELATED INSTRUMENT

A group of objectives which are closely related to those discussed in connection with the discussion of Logical Rea-

²³ This course followed somewhat the pattern outlined by Fawcett, *loc. cit.*

soning and the Nature of Proof relate to what is popularly known as "propaganda analysis." During the Eight-Year Study some attention was given to evaluation with respect to these objectives. This section will give a brief account of this project.

The definition of propaganda which was adopted is as follows: Propaganda represents any use of the spoken or written word, or other forms of symbolization (pictures, movies, plays) designed to convince people to hold certain opinions, to give allegiance to a particular group or cause, or to pursue some kind of social action predetermined by the source of the propaganda. As used in this sense, propaganda has no unpleasant or "bad" overtones. Our concern with it is to better understand which groups are selling what kind of propaganda; the possible social consequences and implications of this; the symbol appeals which are used and their relation to behavior dynamics of individuals; the relation of susceptibility to propaganda to social conditions; etc.

Propaganda also is used to characterize forms of argument which are untenable in terms of certain intellectual or logical criteria such as: documenting evidence, presenting several sides of a problem, drawing conclusions which follow logically from the data, minimizing the use of slogans and "emotional" terms, etc. Used in this sense propaganda does have unpleasant overtones and our problem is to teach pupils to react critically to it by applying criteria of good argument. The scope of this report takes both of these definitions into consideration.

Among the behaviors which were listed as important objectives of education related to propaganda analysis were the following:

- a. Recognition of the purposes of authors of propaganda—that is, ability to make more discriminating judgments as to the points of view which it is intended the consumer

- should accept or reject. (In the broad sense, this refers to the generally accepted concept of "reading comprehension.")
- b. Identification of the forms of argument used in selected statements of propaganda. (This refers to reading comprehension in a different sense.)
 - c. Recognition of forms of argument which are considered intellectually acceptable and which are not employed in certain statements.
 - d. Critical reaction to the forms of argument which represent typical devices employed in propaganda.
 - e. Ability to analyze argument in terms of principles of the nature of proof.
 - f. Recognition of the relation of propaganda to the social forces which breed it.
 - g. Knowledge of the psychological mechanisms involved in the susceptibility of people to certain language symbols.

The evaluation instrument entitled Analysis of Controversial Writing (Form 5.31) was developed to obtain evidence concerning the achievement of the first four behaviors listed above. Item *e* in the list has been discussed at some length above. The others, although they were considered important and some preliminary analyses of them were made, were not explored during the study. The test contains ten samples of writing on controversial issues selected from magazines and newspapers. The choices were made on the basis of the following criteria: (1) the selection should focus upon a controversial issue; (2) liberal and conservative sources were represented on each issue; (3) the group of selections should make use of a variety of propaganda devices; (4) the issues involved should represent areas of tension for pupils.

In each problem the pupils were first directed to read the quotation carefully, and then in Part I to mark them so as to indicate statements where there is:

- A. evidence that the author of the quotation wants you to *agree with or accept* the idea in the statement.
- B. evidence that the author wants you to *disagree with or reject* the idea in the statement.
- C. *no evidence* as to whether the author wants you to agree or disagree with the idea in the statement.

Twelve statements follow these directions. The examples below are taken from Problem I, based on a selection whose tenor may be judged from the closing sentence in one paragraph: "The American system of private industry and business has distributed more income to more people than any other system in the history of the world."

- 1. The present purchasing power of workers is possible only under a system of private ownership of industry.
- 2. Workers should receive higher wages than they receive at present.
- 3. The present system of private ownership is superior to any other way of organizing industry.
- 4. Industry still has far to go in distributing wealth more evenly between the workers and the owners.
- 5. The profits of corporations should be turned over to the workers rather than to stockholders.

In Part II, the student was to decide:²⁴

first, which of the following statements represent forms of arguments used by the author in this situation, and second, which ones represent desirable forms of argument whether used by the author or not.

- 1. Assumes that the point of view expressed in the article is that which is held by the majority of Americans.
- 2. Gives facts in such a way that the reader can check their source to see whether they have been reported accurately.
- 3. Uses statistics for industries in which wages are among the highest to illustrate the rise in wages.

²⁴ The following quotation is an excerpt from the directions.

4. Presents some of the major advantages and disadvantages of our system of private ownership of industry.
5. Indicates that there will be undesirable consequences to industry if our present industrial system is changed.
6. Tries to make us feel sympathetic toward industrial owners.

Ten statements of this general sort were used in Part II of each Problem. In both parts the various statements were so chosen that a student responding according to the directions could reveal evidence of his status with respect to the first four behaviors listed above.

The scores of the pupils in Part I are tabulated in the following descriptive categories:²⁵

General Objectivity. Scores in this category represent the per cent of total correct responses and show the relative objectivity with which the pupil interprets highly biased material.

Non-Recognition of conflicting points of view. Pupils who have difficulty in recognizing ideas which are contradicted by the author's data can be identified through scores in this category.

Misconception of author's purposes. Scores in this category indicate a pupil's tendency to attribute conservative ideas to liberal articles and liberal ideas to conservative articles. Such scores indicate a kind of gross error in judgment and, if relatively large, suggest inability of the pupil to comprehend the general ideas which the authors are trying to sell to the reader.

Suggestibility. Scores in this category indicate the extent to which the pupil indiscriminately attributes conservative ideas to conservative articles and liberal ideas to the liberal articles. (A score of this kind means that the pupil says that the author wants him to "accept" an idea which is keyed "insufficient evidence." The items keyed "insufficient evidence" reflect the general point of view in the articles.)

Except for the category "general objectivity," the scores in Part I categories are separated into "liberal" and "conserv-

²⁵ A more detailed description of how these categories are derived from the test scores and how they are to be interpreted can be found in the "Explanation Sheet and Interpretation Guide" for Form 5.31.

ative." Thus in the "suggestibility" category each pupil has two scores, one showing his suggestibility in interpreting the conservative articles and one showing suggestibility toward the liberal articles.

The scores on Part II are tabulated according to the following categories:

Identification of propaganda techniques used in the articles. This category indicates the degree to which the pupil can recognize the use of the forms of argument keyed as "propaganda techniques."

Confusion of propaganda techniques used and not used. This category shows the extent to which the pupil indicates that the techniques keyed as "not used" were used in the articles.

Uncritical toward the use of propaganda techniques. The tendency of the pupil to approve the use of propaganda techniques is indicated under this heading.

Recognition of acceptable nature of certain forms of argument. Recorded in this category are scores showing whether the pupil approves of the use of the acceptable forms of argument.

Gullibility. Scores in gullibility show the tendency of the pupil to indicate that the acceptable forms of argument keyed as "not used" are used in the articles. Due to the nature of the test items, gullibility means attributing "fairness," "impartiality," "open mindedness" to the authors of the articles.

In constructing Part I of the test, the basic hypothesis was that pupils whose attitudes toward the five social issues included in the test were strongly liberal would tend to be more "suggestible" toward the conservative articles than toward the liberal articles. This was based on the notion that the liberal pupil would more willingly exaggerate the ideas of conservative authors than those of liberal authors. Similarly it was believed that the scores of such pupils in the other columns of Part I would tend to differ as between the sub-categories "liberal" and "conservative." To check this hypothesis an attitude scale consisting of items used in the test was given to approximately one hundred pupils. These

same pupils took Form 5.31 and their attitudes were compared with scores in the "suggestibility" category in the test. This study showed that "liberal" pupils were no more suggestible toward conservative articles than conservative pupils, and vice versa. Furthermore, a study of test scores has shown that most pupils tend to be equally suggestible toward conservative and liberal articles. This same tendency is characteristic of the other categories in Part I.

The conclusion justified from these findings is that the test does not discriminate sharply between the reactions of liberal and conservative pupils in their interpretation of the purposes of the propaganda articles. Sharper differences are discovered when scores on individual articles are compared, for example, scores on the liberal and conservative articles dealing with the issue of socialized medicine. This procedure is cumbersome, however, and would be impractical for use with large classes. Other hypotheses underlying the test seem to be reasonably valid. As one phase of a validity study, 50 essays by pupils who analyzed a subtle piece of propaganda as part of a unit of work on this subject were compared with the test results. The studies of validity and reliability are not complete, however. The instrument has been described because it illustrates an approach to this problem which is somewhat unique and which warrants further study.

CONCLUSION

The two principal uses for these types of instruments are: (1) the diagnosis and description of the strengths and weaknesses of individual students and of groups of students in relation to the objectives as they have been operationally defined in the tests; (2) the measurement of growth in the abilities required for successful achievement. The scores on the data sheets will yield significant descriptions of students with respect to the objectives. The interpreter must, however, clearly understand the structure of the test problems and the relationship of this structure to the problem-solving

process. For certain students the interpreter may desire even more detailed evidence from the test results than that which appears on the data sheet. An examination of the responses of a particular student to certain items on a test may yield such evidence. More often the suggestions raised by an examination of the data sheet will lead the teacher to seek evidence from other sources to confirm or deny these suggestions. For example, a student may reveal a tendency to use many reasons on the nature of proof test but fail to discriminate between relevant and irrelevant reasons. This tendency may or may not be confirmed by the teacher's experience with the student in daily classroom activities.

The uses of these instruments are not fundamentally different from those of many other types of tests. Thus after studying the test results the teacher may wish to provide curriculum experiences designed to overcome obvious weaknesses of a group as a whole, or of individuals within the group. This may lead to a special unit of work for the whole class; special assignments undertaken by a particular student with the advice of the teacher; special attention by the teacher to certain details of the written work handed in by one or more of the students; and the like. In other cases, growth toward this objective might be one of the desired outcomes of the work of a class over a longer period of time. For example, every activity of a class over a period of a year might be designed to make some contribution to the students' concept of proof.

In this connection it will be useful to measure the growth of individuals and of classes toward the objective. Although the students may remember the general nature of these tests for several months, they can scarcely be expected to remember the answers to specific items on the test. Hence the practice effect of taking the tests once will probably not be a serious factor influencing the scores on a second administration of a test several months later. If such studies of growth are desired, it is especially important, of course, that the

specific exercises in the tests should not be "taken up in class." It is also important to keep in mind the effect of the total testing situation upon the test results. This total situation involves more than a careful explanation of the test directions to students, and the provision of adequate time for the completion of the test. In the case of many tests, and particularly those which have been described, it involves also the "readiness" of the class for the test, their attitude toward the test as a diagnostic instrument rather than as a marking device, and the like. Ideally, the class should look upon these tests as an opportunity to demonstrate their ability to do clear thinking rather than as a burden and a threat.

The chief feature of all of these tests is the extent to which they make possible a description of a student's thinking ability in terms of at least tentative answers to a series of questions which are quite general and comprehensive. Successful performance depends relatively little, compared with the usual achievement test, upon knowledge of particular bodies of subject-matter content, and relatively much upon broad principles of science and of scientific thinking. The objectives demanded tests to probe among the higher mental processes applied not to materials of the sort commonly used in psychological investigations, but rather to those commonly found in reading of newspapers and magazines, or elsewhere in daily life. This approach is fundamentally different from one which seeks to synthesize a description of a student's thinking abilities from data on many simpler but more readily controllable psychological reactions. The experience of the Evaluation Staff has been that this endeavor has led to increasing complexity in the test instruments in spite of the demands of practicality for greater simplicity. This increasing complexity was tolerated in order to maintain close correspondence between the stated objectives and the behavior demanded of the student, and in the hope that the instruments of this sort may eventually be simplified.

Chapter III

EVALUATION OF SOCIAL SENSITIVITY



INTRODUCTION

Origin and Scope of the Objectives Related to the Development of Social Sensitivity

In any social situation, an individual is aware of, and responds to, certain factors and lets others go unnoticed. Thus, on observing an old man selling apples on the street corner, one individual may be aware only of the convenience of having apples easily available to him, or be annoyed at having the man clutter up the street corner. The awareness and attendant feelings in this case are self-centered; there is little consideration for the apple man. Another person may "see" primarily an old man trying to make a living. He may in addition feel sympathy for a man who has to make a living in such a precarious way, or feel that this way of earning a living is the man's just due, determined by his ability. Attention in this case is centered on the apple man as a human being. To a third person this experience may suggest the problem of security in old age. He may wonder why there is not a more satisfactory provision for old people to make their living. Awareness and sympathy in this case center not only on the apple man. He becomes a symbol for a whole group of people, or for an issue, and sympathy for him is likely to evoke concern for the problem or issue which he symbolizes. Depending on the type of response, various impulses to action may also suggest themselves. Annoyance with the apple man may suggest activity leading to his removal. Sympathy toward him may lead to consideration of ways of helping

him. Concern about injustice in the social order tends to suggest the need for correcting them.

Several different behaviors are involved in these responses. Personal sympathies and aversions largely determine the pattern of initial awareness. The knowledge one possesses, and the attitudes and viewpoints one has, determine how one interprets the experience. One's ability and inclination to relate and reorganize ideas gained from previous experiences and to apply them to the new situation add insight. The inclination and ability to relate the feelings evoked and positions taken in specific situations to more general and abstract ideas add to both the coherence and the depth of one's insights in a given case. All of these behaviors, although capable of analytic distinction, are related to each other in any given experience.

The term "social sensitivity" has been used to refer to this complex of responses. In the common usage of the term the emotional factors—such as the feelings of sympathy or aversion, attitudes of approval or disapproval—have been emphasized. However, this term can also be used to connote the intellectual responses—such as the range and quality of the elements perceived in a given experience or the significance of the ideas associated with it.

In the first statements of objectives submitted by the schools in the Eight-Year Study the term "social" was used in connection with many types of behavior somewhat similar to the ones described above. Frequent among the statements were terms such as social consciousness, social awareness, social concern, social attitudes, social integration, sense of social responsibility, social understanding, social intelligence. Thus many schools seemed interested in promoting a greater awareness of social aspects of the immediate scene as well as of the issues underlying current social problems. At the same time concern was expressed that unless students achieve clarification of their personal patterns of social values and

beliefs, intelligent social thinking would remain an elusive object of educational effort. The apparent blocking of rational thought by personal prejudices and biases, by a warped sense of values, or by the tendency to react in terms of social stereotypes, was recognized, and many statements of objectives emphasized the importance of a clearer, more consistent, and more objective pattern of social values and beliefs. A good deal of attention was also devoted to the problem of helping students apply the values, loyalties, and beliefs they developed to an increasing range of life problems. The term "social sensitivity" was adopted to serve as a consolidating focus for this apparently heterogeneous yet highly related complex of objectives.

In order to see more clearly what was implied in these statements of objectives from the schools, two committees were established. These committees undertook to make a coherent analysis of social sensitivity as a total objective and to clarify and specify some of the more crucial aspects of it sufficiently to lay a foundation for the development of evaluation instruments. Some of the significant aspects of social sensitivity which were emphasized in the course of the analysis are described in the following section.

Significant Aspects of Social Sensitivity

The first exploratory meetings of the committees revealed a diversity of concepts regarding social sensitivity. In the course of the discussion sensitivity was defined, by implication, as awareness, ways of thinking, interest, attitude, and knowledge. A whole range of problems representing significant areas of social sensitivity was also mentioned. These ranged from such "immediate" social patterns as relations with other people to such general social issues as unemployment, effective democracy, and social justice.

To get a clearer and a more concrete picture of the specific behavior involved in this objective, the committee under-

took to collect anecdotal recordings of behavior incidents illustrating any aspect of social sensitivity which teachers in the Thirty Schools thought important. This material was carefully analyzed and the various types of specific behavior were listed. Altogether, 74 types of behavior were indicated or implied by the anecdotes submitted by committee members and other teachers. The list below gives a few illustrations:

1. The student frequently expresses concern about social problems, issues, and events in conversation, free writing, creative expression, class discussion.
2. The student is fairly well informed on social topics; he has a reasonable background and perspective, and would not often be misled by misstatements.
3. When facing a new situation, problem or idea, he is eager for more information, seeks to identify significant factors in the situation, carries thought beyond the immediate data.
4. He is critical about expressed attitudes and opinions and does not accept them unquestioningly; distinguishes statements of fact from opinion or rumors, discerns motives and prejudices.
5. He is able to discern relevant issues and relationships in problems, ideas, and data. He relates ideas widely and significantly and discriminates among issues.
6. He judges problems and issues in terms of situations, issues, purposes, and consequences involved rather than in terms of fixed, dogmatic precepts, or emotionally wishful thinking.
7. He reads newspapers, magazines, and books on social topics.
8. He is able to formulate a personal point of view; he applies it to an increasingly broader range of issues and problems.

9. He is increasingly consistent in his point of view.
10. He participates effectively in groups concerned with social action.

A classification of these behaviors resulted in the following list of major aspects of social sensitivity of concern to teachers in the Thirty Schools:

1. Social thinking; e.g., the ability (a) to get significant meaning from social facts, (b) to apply social facts and generalizations to new problems, (c) to respond critically and discriminatingly to ideas and arguments. (Statements 4 and 5 above, for example, would fall into this classification.)
2. Social attitudes, beliefs, and values; e.g., the basic personal positions, feelings, and concerns toward social phenomena, institutions, and issues. (Statements 8 and 9.)
3. Social awareness; that is, the range and quality of factors or elements perceived in a situation. (Statements 1 and 6.)
4. Social interests as revealed by liking to engage in socially significant activities. (Statements 3, 7, and 10.)
5. Social information; that is, familiarity with facts and generalizations relevant to significant social problems. (Statements 2 and 3.)
6. Skill in social action, involving familiarity with the techniques of social action as well as ability to use them. (Statement 10.)

The committee on social sensitivity took the responsibility for developing instruments for evaluating three of these aspects; namely, the ability to apply social generalizations and facts to social problems, social attitudes, and social awareness. The present chapter is chiefly devoted to a description

of the instruments pertaining to these aspects. Instruments dealing with other phases of social thinking—such as the interpretation of social data, and critical reactions to arguments and propaganda—have been discussed in the chapter on Aspects of Thinking. The appraisal of social interests is discussed in the chapter on Interests. No new instruments were developed to evaluate the acquisition of social information, primarily because published tests were already available and because teachers felt relatively little need of assistance in this task. As far as securing evidence of skill in social action is concerned, observational records seemed to be the most effective method. These are discussed briefly in the following section.

INFORMAL METHODS OF GETTING EVIDENCE ON SOCIAL SENSITIVITY

An objective which involves as diverse types of behavior as those described in the preceding section obviously necessitates the use of several approaches and several techniques for its appraisal. These will include paper-and-pencil tests as well as observational techniques, each being employed according to its appropriateness to the behavior that is being evaluated. Thus the ability to think through social problems can be adequately appraised by using paper-and-pencil tests. For the evaluation of some other aspects of social sensitivity, such as the identification of social beliefs, paper-and-pencil tests are recommended chiefly because they are economical and because these behaviors are rather difficult to observe directly and objectively. Still other types of behavior, such as the disposition to act on one's beliefs, or the degree of participation in social action and in discussion of social problems, require direct observation of overt behavior. Many of these observational and informal techniques involve only a more effective use of procedures employed and materials secured in the course of normal teaching procedures.

Anecdotal records are an effective way of securing concrete descriptions of significant behavior of individuals or groups. Since they are a way of recording direct observations, anecdotal records are appropriate for securing evidence on all types of overt behavior. However, since such a descriptive record is highly time-consuming, the function of anecdotal records in a comprehensive evaluation program is usually supplementary: to give vivid, intimate, concrete material to help make more meaningful other more systematic but less colorful types of evidence. The nature and role of anecdotes and the criteria for selecting and writing them have been described elsewhere.¹ Here it may suffice to give a few illustrations of anecdotes pertaining to social sensitivity.

A disposition on the part of a group to consider the effects of one's actions upon the welfare of others, and to apply ethical principles in making decisions, is illustrated by the following incident:

The school newspaper had been supported by the income from advertising solicited from small neighborhood stores which the students did not patronize. A student questioned the ethics of such a procedure in the student government assembly. Others in charge of the business management of the paper defended the method on the grounds that it was a general practice with school papers and there was no other way of supporting a printed publication. Another group proposed other ways of earning money, involving more work on the part of the student body. The latter suggestion was accepted.

Class discussion often reveals the degree to which students

¹ The Commission on Secondary School Curriculum, *The Social Studies in General Education*, D. Appleton-Century Company (New York, 1940), pp. 347-50.

L. L. Jarvie and Mark Ellingson, *A Handbook on the Anecdotal Behavior Journal*, University of Chicago Press (Chicago, 1940).

Arthur E. Traxler, *The Nature and Use of Anecdotal Records*, Educational Records Bureau (New York, 1939). Mimeographed.

are capable of using present events to speculate about their consequences.

In connection with a report of a demonstration by members of the League for Industrial Democracy protesting against the "Rex" sailing with munitions for abroad, speculation was aroused regarding the consequences of an embargo. How effective would government control of the sale of munitions be? What devious ways, such as selling to a neutral country, would be devised? (This discussion occurred during the Italian conquest of Ethiopia.)

Personal attitudes toward social issues are often reflected in the daily incidents in the school, as in the following:

Gene came into my room, explaining that she had had an argument with some members of her group over their attitudes during trips they had made to Harlem and the East Side of New York. Jane had told her that she could not see how anybody could like slumming. Gene had objected to such an attitude, since the purpose of the trip was to study the living conditions of people in an unfortunate situation. To her, she said, those trips, together with the study of housing and income, had been one of the most meaningful experiences. She wants to write on that problem.²

Students' writing presents other opportunities for securing evidence on social sensitivity. Much writing contains some expression of social attitudes and of social values held by the author, provided its content is analyzed from that standpoint. Often only a listing of the topics chosen for creative writing over a period of time or for free choice "research" reveals trends in social sensitivity. Thus, frequent choice of social problems to write about or frequent emphasis on social context and social implications is an indication of real interest in social matters. Free choice writing, however, provides

² It is possible to interpret the incidents given above in several different ways. A single incident does not necessarily prove anything about the behavior of an individual and a number of anecdotes covering a period of time must be collected before any generalization is attempted.

only sporadic evidence, and not necessarily on the particular aspects of behavior a teacher may wish to explore. To secure more systematic evidence, controlled assignments in which all students respond to the same general problem, issue, or experience, are often employed. Below is a sample of written responses to the following paragraph assigned as a topic to the whole class: "Nothing can be done about poverty. There have been and always will be poor people, incapable people, unambitious people, dirty work to do, survival of the fittest . . ."

Roy: I think something could be done about poverty. They could be taught many things they have no chance to learn today. They should be housed in a healthy environment. I think there will always be poor people, unambitious people, incapable people, and dirty work to do, but I do not think that a very great percentage of the poor today are poor because of these reasons. They don't have a chance. I don't think that 42 per cent of the Americans today fall into that lazy and unambitious class, yet 42 per cent of Americans are poor. There must be something wrong with our system today.

John: I can find little pity for white and colored trash who have never amounted to anything. . . . I think that the smarter man should make more money and that it would wreck any advancement of civilization so to restrict initiative as to pay the man who carries twice the load as much as the mass below him gets.

Mary: Very few people would at any time . . . be willing to give their money away. Of course, they can be made to give it to the government, but it seems to me to be a shame if people are taxed so heavy to aid all the poor. Surely I agree something could be done, but I can imagine my own feelings if the majority of the voters, who are middle class and poor, should vote for a tax that would take away a large part of the money and savings I had worked for and made.

Even this limited sampling reveals the possibilities of this method of learning about the social viewpoints of the stu-

dents. These excerpts reveal an interesting variety of views regarding causes and cure of poverty and unemployment. Different positions are taken toward taxation. Personal sympathy for people in different economic circumstances or lack of it is shown. One can even gain some idea of the nature and degree of awareness of social conditions in each student.

Records of free choice activities of all sorts often yield surprisingly useful information. Thus records of free reading may give clues regarding students' social interests, level of social awareness, and maturity and direction of social outlook. Records of activities of all sorts, in-school and out-of-school, such as participation in school government, vacation activities, attendance at motion pictures, lectures, and concerts, and other leisure-time activities are also useful, particularly when the nature of the activity is recorded in addition to its frequency.³ Although these records serve primarily as evidence of interests, analysis of their content also serves for evidence of social sensitivity.

Free response tests employing a form akin to projective techniques are also useful devices for getting at personal responses to social issues. Their advantage lies in their indirection. The individual is not asked directly to reveal his social values. He is provided an apparently innocent object of attention to which he can respond freely and personally. The object of attention is so chosen as to draw out revelations of his pattern of social sensitivity. In a completely free response test, only a brief statement is given, and students are asked to list all of the thoughts that occur to them in connection with that subject.

Problem. The following quotation from a local produce market appeared in a daily paper:

"Cooking onions—30 cents per bu."

Directions: List all of your thoughts about this quotation which

³ For further discussion of the use of reading records and activities records, see *The Social Studies in General Education*, pp. 345-46.

might be of social importance. Number your ideas 1, 2, 3, 4, etc.⁴

Certain ideas about students' understanding of, and attitudes evoked by, the problem can be gained from mere examination of each student's responses. However, clearer descriptions of each student, as well as of groups of students, are possible when the responses are summarized in terms of certain general criteria. Thus, the responses to the exercise above could be summarized in terms of the frequency of purely personal association (such as, "I don't like onions"); in terms of frequency of responses showing awareness of the implications of this situation to immediate personal-social values, like the family budget or diet (such as, "If onions are so cheap, they could be used more frequently in family menus"); or, finally, in terms of how frequently the wider social implications are mentioned (such as, "If onions are so cheap, what about the income and the standard of living of those who work in onion fields"). A summary could also be made in terms of how frequently each student mentioned important considerations and how relevant his remarks are to the problem.

More controlled forms of essay tests were used by the evaluation committee in explorations preliminary to the drafting of objective test forms. Students were given a problem situation, with several courses of action listed, and were asked to choose the course of action they thought most desirable. They were then asked to indicate the reasons they would use in supporting their choice.⁵ All such free exercises are, of course, fraught with certain difficulties. To score the responses objectively is a time-consuming process. The fact that each student expresses his thoughts in a somewhat per-

⁴ This exercise was used in Ohio and Michigan at the time when there were strikes in onion fields, and reports of them appeared in the daily newspapers.

⁵ An exercise of this type is discussed on p. 178.

sonal way interferes with the possibility of assigning his responses a precise and fully objective meaning. However, when teachers are able to develop valid exercises of this sort and take the care and the time necessary for a diagnostic analysis of the responses, tests of this sort have a real role to play, particularly since they can be made more readily an integral part of teaching than is the case with more formal tests.

EVALUATION OF THE ABILITY TO APPLY SOCIAL FACTS AND GENERALIZATIONS

The teachers in the Thirty Schools were much concerned that students develop a willingness and ability to use social facts and generalizations, gained through their study, in understanding and explaining social phenomena around them. They recognized the futility of the mastery of a background of facts without growing in ability to apply them to an increasing range of social issues met in daily life. In many schools a serious attempt was made to give students an opportunity to think through new problems in the light of their previous knowledge. For this reason interest was expressed in developing some instruments to appraise students' growth in ability to apply social facts and generalizations.

ANALYSIS OF THE OBJECTIVE

Prior to the development of instruments several explorations seemed necessary. First, it seemed important to identify the generalizations which were considered fundamental to the understanding of social problems and which, therefore, the students were expected to know and to apply in their thinking. It seemed also necessary to analyze and describe the kinds of behavior involved in applying social generalizations and facts. Finally, some exploration of the areas of problems and issues which the students may be expected to be able to think through was also needed. In order to get

some appropriate criteria by which to appraise this aspect of thinking, it seemed important to identify some of the desirable characteristics or qualities of the process of applying social generalizations as well as the difficulties encountered by the students in achieving these qualities. The following sections will discuss these questions in turn and indicate the decisions which were made.

Generalizations and the Processes Involved in Applying Them

Students are often expected to decide whether certain actions—proposed or accomplished—are justifiable, desirable, or reasonable. Such decisions as whether an article attacking democracy submitted to a school paper should be printed, or whether a certain law should be passed in the legislature, are examples. Decisions are presumably made more intelligently when the student understands some of the generalizations which are applicable and has the pertinent facts available. Students may also be expected to explain certain events or to predict the probable consequences. Thus, in predicting the probable effects of a certain type of sales tax, it is important to consider both what is known about the effects of different forms of taxation on various groups in society and certain general principles of taxation. In determining the desirability of the measure in a democratic society, the consideration of certain basic democratic values, such as the welfare of all groups and individuals, and securing equality of sacrifice as far as possible, is also necessary. In much the same way, facts and generalizations are needed in judging the soundness of conclusions drawn or decisions made by other people.

An effort was made by the committee to assemble a representative list of the generalizations taught in social sciences. An initial list was drafted by members of the committee. This list was circulated among the social science

teachers in schools participating in the Study for additions and criticism. Other sources such as Billings' list of social science generalizations and typical textbooks and references were also examined.⁶ The final list was again checked by teachers to indicate which of the generalizations they considered fundamental in understanding social phenomena, which of them were emphasized in their teaching, and which of them were touched upon but not emphasized.

The analysis of this list of generalizations raised several questions about the nature of social science generalizations. In the first place, the line of demarcation between a social fact and a social generalization was not clear. Many of the generalizations listed as major understandings seemed little more than generalized facts and as such had a limited utility in explaining social phenomena other than the ones which they directly summarized. Thus, the generalization that a variety of taxes is levied in the United States adds but little to the understanding of the issues of taxation.

The question of the dependability or the "truth" of many of the generalizations was also raised. Many of these generalizations seemed to apply only to a limited range of situations, and lacked the universality commonly attributed to a "principle," as the term is used in the natural sciences. Often these generalizations seemed little more than hypotheses, useful in exploring ways of explaining events, but questionable for exact prediction. Still other generalizations seemed to have little validity independently of a particular social philosophy or theory. Some generalizations seemed to be direct expressions of the social beliefs held by individual teachers, and the validity of these beliefs was often questioned by other teachers holding different beliefs. It seemed clear that the majority of useful and significant social science

⁶ Neal Billings, *A Determination of Generalizations Basic to the Social Studies Curriculum* (Baltimore, Warwick and York, 1929).

generalizations were not verifiable in the same sense as are the majority of scientific principles.

It seemed advisable, therefore, to think of social science generalizations primarily as tools for further thinking, for formulating tentative explanations, solutions and conclusions, rather than as bases for precise predictions, as infallible guides for action, or as indisputable expressions of "truth." It was finally agreed that the term "social science generalization or principle" would be used to describe any generalization which could be applied to a range of specific situations for the purpose of explanation or prediction, whether or not this generalization was applicable over an indefinitely wide range of such situations or was universally true, precise, or verifiable.⁷

It was clear also that the different types of generalizations suggested involved differences in the ways in which they could be used in the thinking process. On the basis of these differences the principles were classified into three types, each type perhaps implying a different technique for evaluating its use. One group included descriptive generalizations, serving merely to *summarize* a body of discrete facts. Thus, a body of facts about income might be summarized by the generalization, "people earn their incomes through a diversified range of activities." Another type of generalization served to indicate cause-and-effect relationships and to *explain* social phenomena. Thus, a body of data relating to economic penetration into undeveloped countries might be summarized by some such generalization as "economic penetration of an undeveloped country frequently results in military and political domination." A third type had to do with *expressions of value judgments, opinions, or beliefs*. Thus, the body of facts regarding freedom of speech might be

⁷ For a discussion on usage of the term "principle" in curriculum building see Hollis L. Caswell and Doak S. Campbell, *Curriculum Development* (New York, American Book Co., 1935), pp. 87-90.

summarized in the principle, "freedom of speech is essential to the preservation of democracy." This sort of statement expresses a viewpoint or value judgment which is incapable of verification in the usual sense of the term.

The effect of the compilation of such a sample list of generalizations upon teaching was also considered. Some teachers feared that the list would suggest a minimum set of generalizations to be adopted by all teachers and to be taught for memorization. It was agreed the preparation of the list should not be taken to imply that these generalizations had been or should be taught as statements to be learned, but rather that through the best learning procedures the students would be brought to understand certain generalizations, and that they would be given opportunity to apply some of these in their school work. The list was to be used as an illustrative sample of generalizations for the sole purpose of exploring the possibility of evaluating students' ability to apply them.

Analysis of Behavior

In the course of the above discussion some of the behaviors involved in applying facts and principles to social problems have already been indicated.

As was described above, application of principles and facts usually takes place when people are called upon to do any of the following: (a) explain certain ideas or phenomena, (b) predict consequences of events, (c) decide on a course of action, or (d) judge predictions, conclusions, or decisions made by other people. In any of these situations, provided they are new to the students, it is necessary to be aware of the major issues in the problem. A more reasonable judgment usually results when appropriate use is made of whatever facts and generalizations are pertinent to the problem. In the process of making judgments of this type the following behaviors are involved:

1. Relating previously learned facts and generalizations to each other and to the given problem.
2. Discriminating between facts and generalizations which are relevant to a given problem and those which are not.
3. Discerning the logical relationship between a particular conclusion, decision, or a course of action and a generalization or a fact.
4. Organizing facts and principles learned in different contexts in such a way that they can be helpfully used in analyzing the problem or in arriving at the conclusion.

One of the important points brought out in analyzing the objective was that the most fruitful use of important facts and generalizations takes place when these are applied to problems new to the students. Although knowing the facts and generalizations themselves was regarded as basic to the ability to use them, teachers were primarily concerned in this connection with having students develop the ability to organize the facts and principles and relate them to each other in new ways. Hence, the recall of applications made by other people was not considered a behavior to be diagnosed by the prospective instruments.

Criteria for Appraising the Process of Applying Facts and Generalizations

An analysis of the specific behaviors in this type of thinking is helpful, but it is not sufficient for evaluating that behavior. It is also necessary to indicate certain criteria by which to appraise that behavior. Therefore, an attempt was also made to outline the characteristics, both positive and negative, of the process of applying social science principles, which it seemed important and useful to diagnose.

The following characteristics were suggested as important by the committee:

Relevance: Is the student discriminating in his use of generalizations and facts? Are the generalizations which he uses relevant to the situation?

Comprehensiveness: To what extent does the student see the implications of generalizations and facts? What range of important generalizations does he consider? Has he failed to use some of the important generalizations?

Consistency: In the use of value or attitudinal principles does he show consistency in the point of view which he accepts? Does he use some principles which are conflicting either with each other or with the course of action or solution under consideration?

Objectivity and Tenability: Does the student rely primarily on generalizations which can be substantiated by fact, or does he use slogans, emotional phrases, and clichés? Are the statements of facts and generalizations used tenable in the sense that they do not contradict commonly known information?

Selection of Problems

The kinds of problems in which students may be expected to apply facts and generalizations which they have learned were also explored. Again, teachers were asked to submit a list of problem areas dealt with in their classes. A list of 52 problem areas was thus assembled. A considerable range of types of problems was suggested. Some teachers emphasized problems of personal-social relations; others were concerned principally with so-called large social issues. The most frequently mentioned among the latter were: consumer education and advertising, capitalism, distribution of wealth, civil liberties, theories and forms of government, international relations, labor, natural resources, racial issues, profit system, public health, relief, taxes, housing, war and peace, unemployment, public opinion.

CONSTRUCTION OF THE TEST ON THE ABILITY TO APPLY SOCIAL VALUES

The explorations described above determined in several ways the nature of the instruments that were developed. In the first place the analysis of the nature of generalizations indicated that there was a sufficient difference between the processes involved in the application of social values and those involved in the application of non-value generalizations and facts to warrant the use of different evaluation techniques. Accordingly, two instruments were developed: one to deal with the application of value principles or democratic tenets, the other to appraise the application of facts and explanatory generalizations. The first of these instruments, Social Problems (Form 1.41 and Form 1.42), was developed and studied more extensively and is, therefore, reported more completely in this chapter. The second, Application of Social Facts and Generalizations (Form 1.5), is reported briefly.

Several suggestions regarding techniques for the construction of instruments were also derived both from the analysis of the generalizations and of the behavior processes involved in their use. Thus, it seemed to be out of the question to construct exercises requiring students to respond to social generalizations, particularly to value principles, as true and false, or right and wrong. It seemed more appropriate to require students to determine the logical relationships between conclusions, courses of action, and certain generalizations and facts. The very nature of the thinking process in this area indicated that the exercises should take the form of responding to social values in the context of certain problems and issues, and not in isolation. Similarly, the criteria for appraising the process of applying social generalizations, such as relevance, consistency, comprehensiveness, and pattern of values, determined, in a general way, the selection of the

issues to be included in the test, and the sampling of the specific items in the exercises. Thus, in order to appraise the consistency of value pattern it was necessary to include conflicting value principles in each of the exercises. Broadly speaking, then, the categories for the subsequent keying of the test items were determined by a jury of teachers.

Naturally the analysis of the committee suggested only the main structure of the instrument. Additional criteria for the choice and formulation of the items in the test as well as for the choice of summary categories were developed according to what was revealed in the study of the results from the tentative forms of the instrument.

The Choice of the Elements in the Test

In the main it seemed necessary to provide a testing situation in which the students would have an opportunity to take positions or to make decisions about some significant social issues and to support these decisions by using value principles. Consequently the following structure for the test was eventually adopted:

1. A problem situation describing an important issue was presented.
2. Three courses of action representing three different positions toward the issue were formulated. The students were to choose the one or ones which they thought most desirable.
3. A list of "reasons" consisting of value principles was given from which students could choose the ones they would use to support the course of action chosen.
(See illustrative exercise, pp. 180-182.)

As suggested in the analysis of this objective, certain criteria were set up for the choice of the content in each of the three parts of the test mentioned above. Thus, in order to be sure of providing opportunity for applying value principles

and not just *remembering* them, the problems were to be new to the students. Since application to life problems was of concern to teachers, significant contemporary problems were chosen whenever possible; actual problems reported in newspapers or magazines were used. The fact that there are differences of opinion about the value generalizations suggested problems of controversial nature permitting several solutions or conclusions. In order to engage the effort of students, it seemed necessary to select problems which had some significance and meaning to them. Therefore, the tentative formulations of the problems were submitted to students for their criticism and suggestions.

Since solutions to social problems could not be considered as "right" or "wrong" in themselves, the courses of action outlined in the exercises represented different positions and were not to be marked as "right" or "wrong." In order to provide for a diagnosis of different value patterns, it seemed necessary for the courses of action to incorporate the positions currently taken toward the issues described in the problem.

The kind of diagnosis that teachers were interested in making, expressed as criteria for evaluating this type of thinking, suggested the main types of reasons to be included. Thus, in order to discover dominant value patterns, it seemed obvious that statements of contrasting beliefs and values were needed. In order to provide opportunities for students to engage in desirable, as well as undesirable, forms of reasoning, it seemed necessary to include reasons which logically supported each course of action, as well as those which were contradictory, irrelevant, or untenable.

Preliminary Explorations of Test Forms

In order to be sure that the proposed test, in addition to incorporating the desired diagnostic features, would be on a level appropriate to the students who were to take it—that

is would use terms they could understand and include the kinds of values they were familiar with, the types of undesirable reasoning they indulged in, and the kinds of value conflicts current among them—several tentative drafts of the test were tried out.

Ten "direct form" exercises were drafted. Each contained a statement of a problem, and three courses of action. Students were asked to choose from these alternative courses of action or conclusions those which they approved and to write out their own reasons to support their choices.

SAMPLE EXERCISE:

Cotton Picker. Cotton has been picked by hand, which is a slow and expensive process. Recently, the Rust brothers invented a machine to do this work. It would pick in $7\frac{1}{2}$ hours as much cotton as one hand picker could pick over a whole season of eleven weeks. The cost of production of cotton could be reduced from \$14.52 to \$3.00 per bale. To date, this machine has not been placed on the market. What should be done with this machine?

Solutions: (Check one or more which you think are desirable.)

- A. The machine should be placed on the commercial market for immediate manufacture and sale.
- B. The machine should be made available under some form of public control and provisions made for establishing in other jobs the cotton pickers who are thrown out of work.
- C. The machine should not be put to use at the present time.

Directions: Write in the space below the reasons which you would use to support the solution or solutions you have checked. Be sure to write all of the reasons you can think of.

Below is a sample of the reasons used by the students checking the course of action A:

1. The normal trend of business would reemploy the replaced workers gradually.
2. The cotton workers could always go on temporary relief.

3. When a good invention like this has to be withheld from the market because of the problem of what to do with the unemployed, it is a little doubtful whether our present economic system is really serviceable.
4. Society should not be deprived of anything that might improve work and the products it uses.
5. Economic statistics prove that there is no such thing as technological unemployment.

These student responses were used in several ways in drafting the instrument. In the first place, it was possible to check the usefulness and the validity of the criteria for summarizing and evaluating the responses suggested by the committee. It was found that most of them—comprehensiveness, consistency, relevance, tenability, and patterns of values—were useful in classifying and summarizing student responses. Thus variations were found in the range of implications seen (comprehensiveness). Often the reasons chosen by the students were in conflict with the courses of action they had marked (inconsistencies). Many students used reasons contrary to facts (untenable) or which did not apply to the courses of action they had chosen (irrelevant). Different value patterns were also expressed. These value patterns were at first summarized under the following headings: protection of human values, consideration of general public welfare, democratic tenets, desire for justice, approval of change, protection of the economic interests of property owners, protection of the interests of privileged groups, economic individualism, safeguarding of present institutions, laws, and customs. Because of the limitation in the length of the test, later it was necessary to reduce this classification to the following one: democratic values, undemocratic values, and rationalizations. In the second place, student responses also suggested the content for each variety of reasons. Thus, the types of untenable and irrelevant reasons to be used, the kinds of inconsistencies, and the kinds of democratic and

undemocratic values to be included were largely determined by analysis of these free responses. Suggestions were also found regarding terminology suitable for use in the test statements. The final form of the test included many statements made by the students. In other cases the phrasing as well as the content was patterned closely to the statements made by the students.

Description of the Final Test

A sample of the final test exercise with an example of some of the reasons is given below. The key is inscribed on the margin.

PROBLEM IV. "WORKING CONDITIONS"

Each year many workers have to stop working either temporarily or permanently because they develop poor lung conditions, arthritis, rheumatism, or just general ill health. It is known that such factors as dust, dampness, and unregulated temperature greatly contribute to these ailments, though it is impossible to determine in many individual cases to what extent the illness was caused by these conditions.

Since it would involve costly improvements to eliminate these conditions, many mines and factories have done little about them and oppose further regulation. With the exception of a few states which have adequate health regulations, at present only such things as hours of work and conditions leading to accidents are regulated by the government.

What should be done about such problems?

Directions: Choose the most acceptable course (or courses) of action and fill in the appropriate spaces on the answer sheet under Problem IV.

Courses of Action:

- | | |
|----------------|--|
| (Undemocratic) | A. It should be left to the individual mine and factory owners to determine what is needed and what they can afford. |
| (Democratic) | B. Minimum standards for general working |

conditions, including all factors injurious to health, should be set by the government and all industries should be required to meet these standards.

- (Compromise) C. In industries where such conditions are likely to prevail, improvements should be made on the basis of suggestions from joint committees of workers and employers.

What reasons would you use to support your course (or courses) of action?

Directions: Choose the reasons which are in harmony with what you believe and which you would use to support your course (or courses) of action and fill the spaces on the answer sheet in the column under the course of action you marked at the top. If you have chosen more than one course of action, and a reason supports both, mark it in both columns.

Reasons:

Key

Supports A and C

Inconsistent with B

Rationalization

Supports A and C

Inconsistent with B

Undemocratic Value

Supports C

Inconsistent with

A and B. Democratic

Value

Supports B

Inconsistent with A

Irrelevant to C

Democratic Value

Supports A and C

Inconsistent with B

Undemocratic Value

1. It would be unfair to require factories to introduce costly improvements which they feel they cannot afford.
2. Without regulation, business can be depended upon to make necessary improvements.
6. If workers participate successfully in solving this problem, there is likely to be further cooperation between employers and employees.
8. Human welfare should be protected regardless of the cost to industry.
10. Since employers have to bear the expense of making improvements in working conditions, they should have

Key

- | | |
|---|--|
| | a voice in deciding what changes should be made. |
| Untenable | 12. Most industries today provide as healthy working conditions as they can afford without undue strain on their finances. |
| Supports A | 15. If a worker is willing to accept employment in an industry, he should be willing to work under the conditions prevailing in that industry. |
| Inconsistent with B and C. Undemocratic Value | 16. Even though it is important to improve working conditions, it is undemocratic to accomplish this through dictation by the government. |
| Supports A and C | |
| Inconsistent with B | 20. In the past improvements in working conditions have come only under government compulsion. |
| Rationalization | |
| Untenable | |

A word of explanation may be necessary regarding the method of arriving at the key for this instrument. The analysis made by the committee suggested the classification of all items, except the specific diagnosis of the value pattern. This was developed by an analysis of responses and was checked by teachers. The items were keyed by a jury composed of members of the Evaluation Staff and some teachers of social sciences.

On the assumption that value preferences and logical judgments both enter into and influence each other in the normal life response to controversial social issues, evaluation processes should not isolate these behaviors and treat them as if they occurred independently of each other. Hence, the test is not made up of parts corresponding to each of the aspects of behavior measured by the test.

Only one process of marking the test is employed: the students mark the reasons which they would use to support the courses of action they chose. The students use each reason

only once with each course of action. But each reason is keyed in several different ways. Thus, reason 1 in the above exercise supports courses of action A and C and is inconsistent with B. Depending on the course of action with which it is used, response to this reason is scored under the accurate reasons contributing to comprehensiveness, or under inconsistency. In addition, each exercise contains two or three reasons which are contrary to commonly known facts, i.e., are untenable (reasons 12, 20). These reasons are not keyed to any particular course of action, but are so sampled that for each position there is one untenable reason which has some logical relationship to it. They are scored as untenable no matter with which course of action they are used.

Most of the reasons are also keyed to represent value positions, as are all courses of action. The value patterns are grouped into three categories: (1) *democratic*,⁸ representing defense of the interests of the general public or general welfare, of such democratic rights as freedom of speech, equality of opportunity, and a decent standard of living, of rights of minorities and other underprivileged groups (course of action B, reasons 2, 6, 8); (2) *undemocratic*, representing protection of special privilege, supremacy of efficiency and economic gain over human needs and values, undemocratic procedures, or discrimination (course of action A, reasons 10, 15); (3) *compromise*, representing essentially an effort to reconcile these two types of values (only courses of action, e.g. C, are used). *Rationalizations* (reasons 1, 16), representing undemocratic values stated as democratic slogans, are keyed and scored as a separate value category, but they can be used to support either the undemocratic or compromise courses of action. At least six supporting reasons are available for each course of action. The logically sound sup-

⁸ The meaning of the terms "democratic" and "undemocratic" as used in this test have thus a special definition, somewhat more encompassing than the common usage of these terms.

port for the democratic course of action is composed exclusively of democratic values. Those supporting the undemocratic course of action are all undemocratic values. About half of the supporting reasons for the compromise course of action are democratic and half are undemocratic values. No matter with which course of action the reason is used, it is keyed to the same value. Thus reason 1 is keyed as a democratic value and reason 2 as an undemocratic value, independently of the course of action with which they are used.

In the entire test there are eight of these exercises, covering such problems as conservation of national resources, free speech, unemployment, protection of health, distribution of wealth, collective bargaining, and socialized medicine. The pattern of reasons described above is the same in all eight exercises.

Summarizing and Interpreting the Results

On the sample form of a data sheet shown on page 185 the scores for four students are presented for purposes of illustration. At the bottom of the data sheet the maximum possible score, the highest score, lowest score, and the group median are recorded for each column. All of these are computed for the class of 53 students from which these four were drawn.

Scores on this test can be interpreted in terms of answers to three questions. The first of these questions is: *How broadly does the pupil relate principles or value generalizations to chosen courses of action?*

Comprehensiveness (columns 1, 2, 3, 4). The most important score here is found in the column headed *Ratio* (column 4). This score is the average number of logically accurate reasons the student has marked for each course of action. A high score here usually indicates the ability to see the implications of social values in concrete social problems broadly and fully. Thus Student A has one of the highest scores in the whole class on comprehensiveness (6.1), while Student

Score	Comprehensiveness				Logical Discrimination			Dominant Values						
	Courses of Action	Total Reas.	Accurate Reas.	Ratio	% Inconsistency	Un-ten-able	Irrel- evant	Courses of Action			Reasons			% Demo- cratic Values
								Dem- ocratic	Un- demo- cratic	Com- pro- mise	Dem- ocratic	Un- demo- cratic	Ra- tional- iza- tions	
Column Number →	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Student: A B C D	9	71	55	6.1	8	4	6	5	0	4	3	43	8	81
	8	41	19	2.4	28	7	4	1	2	5	13	6	8	22
	8	59	38	4.8	10	8	6	1	1	6	10	15	13	42
	8	21	18	2.3	0	2	1	6	0	2	0	14	0	100
Maximum Possible	8	154	D53 U53 C55			D9 U12 C14	D8 U9 C8	8	8	8	57	56	26	100
Lowest Score	5	21	17	1.6	0	1	0	0	0	0	0	6	0	20
Highest Score	12	87	64	7.9	44	11	10	8	6	6	15	53	15	100
Median	8	54	39	5.0	5	5	3	4	0	3	4	26	7	58

D has used on the average only 2.3 reasons for each course of action that he chose—a ratio score which is considerably below the median. This suggests that Student A has a much broader vision of the implications of social values than does Student D. The scores on *total* reasons (column 2) and *accurate* reasons (column 3) are for purposes of reference only. Thus occasionally it is important to see whether a student has marked many reasons in excess of those needed to support his position. This would suggest that the student is confused or lacks discrimination, which, for instance, is the case with Students B and C. Each has used over 20 reasons which do not support the courses of action he chose. In the case of Student B, these constitute over half of the total reasons marked.

The second question is: *To what extent does the pupil show lack of logical discrimination in the use of reasons to support the courses of action which he chooses?*

Undesirable Reasons (columns 5, 6, 7).

Per Cent Inconsistency (column 5). This score gives the per cent of the total number of reasons checked by the student which are inconsistent with the course of action chosen. A high score here indicates inability to see clearly the logical relations between value principles and social issues. As such, it is an index either of lack of ability to deal with abstract principles or else of a confused value pattern which makes it impossible to see their implications clearly. Student D has avoided all inconsistencies, while 28 per cent of the reasons marked by Student B were inconsistent with the courses of action chosen, the median for the class for inconsistency being 5.

Untenable Reasons (column 6). This score gives the number of reasons checked by the student which are contrary to commonly known facts. A high score here indicates either a tendency to use questionable evidence to support one's position, or it expresses idealistic naïveté and goodwill toward

social conditions and a lack of awareness of the real conditions. Student C uses eight such reasons, while Student D uses only two. It must be observed, however, that the range for this score is small.

Irrelevant Reasons (column 7). This score gives the number of reasons checked by the student which do not apply to the particular course of action chosen. A high score here suggests lack of discrimination between reasons that are relevant and those which do not apply to a given course of action. Students A and C show higher than average tendency to fail to discriminate between the relevant and irrelevant reasons, while Student D has marked only one irrelevant reason.

The third of these questions is: *What values are dominant among the courses of action and reasons chosen by the student?*

While the choices of courses of action as well as of reasons yield information on patterns of value, the former are used only in a subsidiary fashion to determine the consistency of the pattern. The scores on reasons (columns 11 to 14) are of primary importance here. Those on courses of action can be used only as supplementary evidence. The main score on dominant values is the *per cent democratic values* (column 14). A high score here indicates a clear-cut and exclusive acceptance of the democratic values as defined above (p. 183). One hundred per cent of the values used by Student D are *democratic*, while only 22 per cent of the value reasons used by Student B fall in this category.

Columns 11, 12, and 13 represent a more specific analysis of the distribution of value scores.

Democratic. Scores in column 11 report the number of times the student has used reasons expressing the values of general welfare and democratic rights. Student A uses a large number (43) of values of this type, while Student B has a score of 6, which is at the bottom of distribution for the class.

Undemocratic. Scores in column 12 give the number of

reasons which express defense of the interests of special privilege of all sorts. A high score here indicates a predominant acceptance of *undemocratic* viewpoints on social issues, as defined above (p. 183). Student B has used 13 of this type of value statements. This is not only considerably above the median but also this type of value composes the largest part of the total value reasons used by him.

Rationalization. (Scores in column 13). Included under this heading are reasons which attempt to rationalize an essentially *undemocratic* viewpoint by couching it in democratic terminology. High scores here indicate a tendency of gullibility to slogans and an inclination to pay lip service to democratic generalities. Student C shows such an inclination, having used more than the average of these reasons.

Sometimes it is worth while to compare the values expressed in choices of courses of action with the value pattern revealed in reasons. Often these two aspects of reasoning are not consistent with each other. Thus if the majority of the reasons checked by the student are democratic values but several undemocratic or compromise courses of action are chosen at the same time, one may infer that the student does not fully see the implications of the values he accepts verbally. Such seems to be the case with Student D. He has chosen two compromise courses of action, which normally call for part democratic, part undemocratic support, yet he has used no undemocratic values among his supporting reasons.

In the foregoing explanation each of the scores was considered independently. This is normally the first step in interpretation. Since each of the single scores describes only one part of a pattern, it is not justifiable to draw conclusions about an individual without considering the whole pattern of scores. In such a pattern, a score often assumes a meaning which differs from the one gained from considering it by itself. In attempting a pattern interpretation it is useful to

consider these scores in two groups: one representing the logical aspects (comprehensiveness, consistency, tenability, relevance; columns 1 to 7), the other representing the pattern of values (democratic, undemocratic, rationalization; columns 8 to 14). However, in addition to examining the scores in each major group in relation to each other, it is also necessary to consider the logical aspects in the light of the value pattern and vice versa.

Student A, for instance, tends to be comprehensive in his use of reasons. At the same time he is somewhat lacking in logical discrimination, as shown by his tendency to accept inconsistent and irrelevant statements in supporting the courses of action. Since his dominant value pattern is democratic in a clear-cut way, one is led to infer that his main difficulty is weakness in logical discrimination.

Student C shows confusion both in the logical aspects (relatively high inconsistency) and in his value pattern as shown by his frequent choice of compromise courses of action and of rationalizations. One might infer from this that his difficulties with logical aspects of applying values stem from the confusion of the values he accepts. His scores on democratic and undemocratic values are rather evenly divided and a high score on rationalizations suggests gullibility to democratic slogans. When one considers in addition the fact that he uses only a few supporting reasons, one is forced to describe the whole picture as that of a lack of awareness of, and confusion about, social issues.

A high degree of inconsistency is one of the major facts about Student B. But because his value pattern tends rather clearly toward the undemocratic one, one is forced to conclude that his main difficulty is that of misapprehension of logical relationships between reasons and courses of action.

The patterns of reasoning illustrated above are found recurrently among students. Some students may be broad in their reasoning and at the same time consistent, discriminat-

ing, and have a clear value pattern. Others may be broad, but inconsistent and ambivalent in value pattern. Some are narrow, clear, and have a democratic value pattern. Others may be ambivalent in their values, but not inconsistent. This usually happens when they take different positions regarding the different issues included in the test but are not confused as far as the same issue is concerned. For teachers interested in diagnosis of the kinds of thinking students do and of the ways their value patterns either help or hinder that thinking, this is useful information.

VALIDITY AND RELIABILITY

The usefulness of this instrument, as of any instrument, is determined by (1) how adequately it measures what it sets out to measure (validity) and (2) how reliable a particular set of the students' responses is likely to be. The problem of validity is a complex one and includes the consideration of the validity of the instrument itself, as well as of the conditions under which the test is given and taken. In this section attention is devoted to the discussion of the validity of the instrument itself. The conditions under which valid results are possible in a given situation will be discussed in the section on uses.

The validity of the results from a test of this type is determined by several factors. In the first place, there may be a difference between the behavior specified in analysis and the behavior actually measured by the test. Any test situation is an artificial situation and may introduce difficulties irrelevant to its purpose. Hence, it is important to see what correspondence there is between the evidence from the test and that obtained from freer and more natural situations.

Each test also employs a certain method of scoring and summarizing. This method may not give the most adequate picture of the responses to the test and therefore it is neces-

sary to determine how effective the method of scoring and summarizing is.

Finally, there is always the question of the degree to which general ability affects success with a given test. This test does not purport to be a measure of general intelligence. Therefore, some evidence is needed to determine the relation of this factor to the responses to this instrument.

Some evidence was secured on all of these points in the course of the study. Serious effort was made in the process of constructing the test to assure as great a degree of validity as possible. Throughout the process of construction steps were taken to make sure that the test appraised the behaviors it was intended to appraise. As was indicated in the description of the preliminary analysis and of the exploratory studies, care was taken to see to it that the behavior measured as well as the content of the exercises was appropriate to the students who were to be tested and consistent with the objectives and curriculum emphasis of the schools. The problems and generalizations included in the test were chosen according to what was found to be most widely emphasized in the schools intending to use the test. Student responses to essay forms were examined to secure reasons representing the types of values and patterns of reasoning current among the students. In addition, tentative drafts of the more objective forms were tried out and revisions were made on the basis of the responses.

Similar explorations were conducted to develop the most useful categories of summary and methods of scoring. The initial choice of the summary categories was made according to the suggestions made by the committee. These were tried out experimentally, and revisions and additions were made according to the dependability and usefulness of the particular scores as shown by experimental use of the test. Thus, for instance, some of the rather fine classifications of values attempted at first proved impracticable because the test

could not be made long enough to get high reliability on these scores.

The validity of the diagnostic descriptions of students made from the test scores was also checked informally throughout the Study. In each school where the test was given conferences were held with the faculty. Students selected by the faculty were described on the basis of the test scores and these descriptions were submitted to the collective judgment of the faculty. Usually students who were known by most teachers, and intimately known by some, were chosen for this purpose. This was done in about 25 of the Thirty Schools, and descriptions of several hundred students were thus examined and checked in the course of two years. Outright disagreements on major points were rare. These occurred mostly in cases where the observations of different teachers varied considerably.

Certain difficulties were experienced in the use of the usual statistical techniques for estimating validity and reliability. The scores describing the logical aspects and those describing the value judgments are both derived from a single process of marking by the student. Each aspect influences the other, however, and interpretation must account for this interrelationship. Thus a high score on comprehensiveness combined with high consistency means one thing. The same score on comprehensiveness combined with high inconsistency means something different.

However, statistical techniques which are simple enough for practical purposes in an exploratory study such as this one do not permit the treatment of the validity and reliability data in terms of a pattern of scores. They usually are predicated on the assumption that each score is a separate entity. Hence it is felt that the quantitative evidence presented in substantiating the claims for a certain degree of validity and reliability of the instrument do not do full justice to it.

Validity was investigated by the following three methods:

(1) comparison of teacher observations with test scores, (2) comparison of interviews with students with the test materials, (3) correlation of the scores on this test with scores on psychological tests.

The comparison of teacher observations with the test results was employed with the full recognition of the fact that the opportunities for teachers to observe these particular characteristics were apt to be deficient and hence not fully reliable. In three schools a selected group of teachers was asked to rate a group of senior students separately on the three major characteristics diagnosed in the test: comprehensiveness in seeing implications of social values, consistency of their social reasoning, and the pattern of social values. Altogether, 132 students from three schools were thus rated. From five to eight teachers in each school participated, with an average of four teachers rating each student. A three-point scale (1—high, 2—average, 3—low) was used for each of the characteristics. These ratings were then compared with the corresponding test ratings. The results are presented in the table below.

MEAN SQUARE CONTINGENCY CORRELATIONS
OF TEACHER RATINGS AND TEST RATINGS

	Compre- hensiveness	Consistency	Democratic Values
School I	.49	.63	.38
School II	.50	.66	.64
School III	.60	.41	.58
One teacher in School II	.78	.79	.88

These data suggest that, on the whole, there is a general agreement between teachers' ratings and test ratings. All correlations are positive and with three exceptions are .50 or higher. The highest relationships were found in School II, in which the teachers participating in the rating had the best

opportunities to observe their students. The ratings of the student adviser in the same school have the highest correspondence with test scores. Thus the relationship between the test and the teacher ratings seems to increase as the conditions necessary for reliable teacher rating improve. This would suggest that the reliability of teacher ratings is a strong factor in limiting the correspondence. It should also be remembered that while in the normal process of interpreting the results of this test the meaning of a single score is often altered in the light of the whole pattern of scores, single scores were used in the statistical processes of computing the correlations. Hence, the coefficients expressing the correspondence are apt to be lower than would have been the case had it been possible to use all scores in relationship to each other. However, in spite of these difficulties, these data suggest that when thoughtful judgments are made by teachers who have had adequate opportunity to observe students' social thinking, a rather close agreement is likely to occur. These data are also in accord with the hypothesis that under usual classroom conditions teachers would be able to identify most of the extreme cases without the test, but that close agreement throughout between the test and teacher rating would not be found, since teachers ordinarily do not have a very adequate basis for observing these particular qualities and hence for rating them very precisely.

Another method used was that of interviewing the students. Forty-five students, 15 from each of three schools, were interviewed. Their specific responses to the test items were first analyzed and summarized in a written statement. The students were then interviewed regarding their viewpoints on social issues included in the test. Through a series of questions, the students were led to comment on the kinds of solutions they approved and the reasons why they thought these solutions were appropriate. Verbatim records of these interviews were taken. The itemized analysis of the test re-

sponses and interview records were then submitted to four judges, all of whom were familiar with what the test was attempting to measure. These judges were first asked to indicate the extent of agreement between what the students said in the interview and how they had marked each exercise in the test. This agreement was rated on a three-point scale: 1—good, 2—fair, 3—poor. An average rating for the degree of agreement for each student throughout the test was compounded by adding the values of all judges' ratings on all exercises and by dividing this total by the number of ratings.

In most cases the agreement was found to be high. Thus, the mean rating on all students on all problems was 1.29, indicating only slightly less than "good" correspondence in the majority of cases. The lowest average rating on any student was slightly better than "fair" (1.78). The number of "good" ratings represented 75 per cent of the total number of ratings, while the number of cases of poor correspondence represented 3 per cent of the total ratings. Thus it is apparent that these judges considered the interview materials to be highly consistent with the test responses. This is particularly gratifying in view of the fact that several students confessed a change of viewpoint between the taking of the test and the interview.

Three of the judges were then asked to consider the interview materials alone and to rate each student on three aspects measured in the test: comprehensiveness, consistency, and pattern of values, on a three-point scale (high, average, low), in order to get some evidence of the adequacy of the summarization and scoring. These ratings were correlated with the test ratings on the corresponding scores, with the following results (expressed as product-moment correlations): comprehensiveness .59, consistency .51, democratic value .66. Considering the meagerness of the interview materials for rating purposes and the fact that the interviews were conducted on topics similar to but not the same as the

test exercises, and taking account of the difficulty involved in treating the test scores in isolation from each other, it is justifiable to assume that the method of scoring and summarizing represents student responses to the test fairly adequately.

In order to see to what degree general intelligence is related to the results on this test, the scores on the American Council Psychological Examination for 45 students were correlated with the three main scores on this test. The relationship was found to be low on all three; namely, comprehensiveness .27, consistency .35, democratic values .04. The number of students is too small to afford conclusive evidence, but there is a fair indication that the performance on this test is relatively independent of the abilities measured by the psychological examination.

Several checks were also made of various aspects of reliability. The stability of scores was tested by several methods of estimating reliability. The split-half method was used on scores which permitted such treatment. The Kuder-Richardson formula was used wherever the split-half method did not apply.⁹ The estimated reliability for the score on per cent democratic values was obtained by correlating Forms 1.41 and 1.42 of the test. The coefficients of correlation secured from a sample of 600 students in tenth, eleventh, and twelfth grades range from .50 (untenable) to .91 (democratic values).¹⁰

On the chief scores used in interpreting the results (comprehensiveness ratio, per cent inconsistency, number democratic values, number undemocratic values, per cent democratic values), the reliabilities range from .70 to .91, which may be considered fairly high for a test of this type, particularly since the final judgment of the students' behavior is based on a pattern of scores and does not depend exclusively

⁹ *Loc. cit.*

¹⁰ See Appendix for a complete table of reliability coefficients by grades.

on any one single score. Low reliabilities were found on the scores on untenable reasons (.50) and rationalizations (.67).

These data seem to indicate that this test has sufficient validity and reliability to be a useful instrument for diagnosis. It must be remembered that the behavior measured in this test is highly complex, affected by variability in the interpretation of test statements and by emotionalized responses. Hence, objective tests in this area probably cannot be judged by the same criteria as are applied, for instance, to tests measuring achievement in acquiring information. It is also likely that under optimum conditions, where teachers have worked seriously on this objective, and students are familiar with the type of reasoning and the kind of content involved, both the reliability and validity estimates might be higher.

APPLYING SOCIAL FACTS AND GENERALIZATIONS TO SOCIAL PROBLEMS (FORM 1.5)

As was pointed out above, teachers of the social studies were concerned with students' ability to apply not only value judgments but also relevant and accurate *information* in their analysis of social problems. An instrument developed to get evidence of the latter ability will be described briefly, since the processes involved in its construction were analogous to those reported at length in the preceding section.

Analysis of the Objective

The analysis of the objective resulted in the following list of important types of behavior to be evaluated: (1) *The ability to see the logical relations* between general principles and specific information on the one hand and the issues involved in a given social problem on the other; i.e., to see whether a statement supports, contradicts, or is irrelevant to a conclusion. (2) *The ability to evaluate arguments* presented in discussing a specific social problem, and in par-

ticular, to discriminate between statements of verifiable fact, statements of opinion and common misconceptions. (3) *The ability to judge the consistency of social policies with social goals*; i.e., to judge the appropriateness of certain social policies for achieving certain social aims.

There are two major types of situations in which individuals make use of these abilities: (1) when one evaluates a proposed solution of any social problem, and (2) when one proposes a solution and tries to support it. The test described below is based upon the first type. These situations occur in the consideration of a wide variety of problems, involving many types of generalizations and of factual information. Before any instruments could be developed in this field, it was necessary to make a choice of problem areas and types of generalizations to be sampled. The list of social science generalizations and of significant problem areas submitted by the teachers and discussed above was used as the primary source of issues upon which to build the test.¹¹ These were checked further with respect to the frequency with which they occurred in high school courses on social problems. The following problem areas were selected: consumer buying, health, unemployment, housing, soil conservation, civil liberties, international relations, taxation, and civil service.

Description of the Instrument

Exercises were constructed for each of the problem areas listed above. Each exercise is a complete test in itself and can be used independently of the others. An exercise is composed of several parts, constructed in such a way as to give evidence of the three abilities listed in the analysis of the objective. In the first part of the exercise a social problem is described, and one of the frequently suggested solutions is indicated. Various statements (some supporting, some contradicting, and some irrelevant) concerning the solution are

¹¹ See p. 170

presented. The student is asked to indicate whether each statement supports, contradicts, or is irrelevant to the suggested solution. A student's reactions to this part of the test are summarized in terms of the number of accurate responses he makes, the number of times he confuses supporting and contradictory statements, and the number of times he fails to see the relevance of a statement to the conclusion. The statements include basic assumptions, general principles, accurate information, and common misconceptions. In the second part of the test the student is asked to indicate whether each of the statements can be proved to be either true or false. The student's reactions to this section are summarized in terms of the number of times he discriminates between statements of fact and assumptions, the number of times he marks value judgments as verifiable, the number of times he marks statements of fact as not verifiable, and the number of times he discriminates accurately between true statements and common misconceptions. An excerpt from one exercise is given below. The key is indicated at the left of each statement.

HOUSING¹²*Application of
Principles 1.5*Form I
(Tentative Draft)*Problem:*

Housing is one of the problems of concern today. Many schemes have been suggested as a means of improving housing conditions. In general, there are two major ways in which government can aid in solving this problem: (1) by setting standards for and regulating the construction of private housing, and (2) by building houses at public expense, contributing either part or all of the funds necessary. Each method has certain advantages and disadvantages. Nevertheless, many people believe that the government should build houses at public expense to rent to those sections of the population with the lowest incomes.

¹² In all cases where the phrase "decent house" or its equivalent is used, it is to be defined as a separate house or apartment for each family with running water, inside bath, fire protection, and enough room for privacy.

I. Directions: For each of the following statements, place a check mark (✓) in one of the columns labeled Part I. Place the check mark (✓) opposite the number which corresponds to the number of the statement in:

Column A if the statement may logically be used to support the underlined conclusion.

Column B if the statement may logically be used to contradict the underlined conclusion.

Column C if the statement neither supports nor contradicts the underlined conclusion.

Check each item in only one column. In case of doubt, give the answer which seems most nearly right.

In this part of the exercise, assume that each statement is true.

Supports Assumption	1. Whenever houses are not available to the public, society should assume the responsibility for making it possible for everyone to have a decent place to live.
Contradicts Misconception	3. Government-built houses are more expensive to construct than comparable houses built by private companies.
Supports Misconception	11. It has been demonstrated that the federal government can build adequate houses for the lowest income group cheaply enough so that they can be paid for out of income from rent.
Contradicts Accurate	14. Individuals who have heavy investments in slum property would probably suffer heavy losses if a broad program of federal housing went into effect.
Contradicts Assumption	17. The system of private initiative in business should not be jeopardized by the socialization of any of the fundamental industries.
Supports Accurate	20. Under present conditions, at least 50 per cent of the people cannot easily afford to own a decent home; at least one-third of the population cannot afford to rent decent homes.

Irrelevant
Accurate

22. Comparable houses can frequently be rented in the suburbs for somewhat lower rentals than in the city.

II. Directions: Go back over the statements. In the columns labeled Part II place a check mark (✓) opposite the number which corresponds to the number of the statement in:

Column D if you believe that the statement can be proved to be true.

Column E if you believe that the statement can be proved to be false.

Column F if you believe the statement cannot be proved to be either true or false.

Check each item in only one column. In case of doubt, give what seems to you to be the one best answer.

When you have finished Part II, go on to Part III.

A student may be able to make the logical analysis and to evaluate the argument very accurately and yet may not be able to judge whether or not a given social policy is likely to achieve a given social objective. Therefore, in the third part of the test the student is given opportunity to make this type of judgment. This part of the test consists of a statement of a particular social objective (such as the improvement of the housing conditions of the third of the population with the lowest income), and several proposals, some appropriate, some inappropriate, for achieving this objective. The student is asked to indicate which proposals he thinks would be effective in achieving the objective. His reactions to this section of the test are summarized in terms of the number of times he chooses policies which are helpful in achieving the stated objective.

An illustration of this part of the test is given below.

III. Directions: In the column labeled Part III opposite the number which corresponds to the number of the statement, write:

A plus sign (+) if it expresses a type of action which you think would improve the housing conditions of that third of the population with the lowest incomes.

A zero sign (0) if it does not express a type of action which you think would improve the housing conditions of that third of the population with the lowest incomes.

- + 1. New buildings should be required to measure up to higher minimum standards for construction.
- + 2. Credit for housing should be supplied in larger quantities and at lower rates of interest.
- 0 3. All city land should be reassessed.
- 0 4. Laws should be passed requiring the destruction of all slum areas.
- + 5. The government should subsidize housing for lower income groups.

Accurate response to each of the first three steps involves the use of certain general information. In case the student makes a large number of inaccurate responses, it is important to know whether it is because he does not have the information or whether he knows the facts of the situation but cannot apply them. Therefore, in the last section of the test the student is asked to judge the truth or falsity of a series of statements which sample the information that is assumed in the preceding sections of the test.

A sample of the factual statements in this section of the test which correspond to the arguments used in the illustration of Part I is given below:

Directions: Form II. The following items refer to the problem of housing. In the columns labeled Form II place a check mark (✓) opposite the number which corresponds to the number of the statement, in:

Column X if you believe the statement to be true.

Column Y if you believe the statement to be false.

Column Z if you are uncertain whether the statement is true or false.

- True 1. At present various estimates agree that at least one-third of the population lives in unsanitary or unhealthy homes.
- False 3. On the average, the cost of federal housing has been approximately \$1,000 more per unit than the cost of comparable private construction.
- False 11. To date the income from rent on housing projects has been large enough to pay for the original cost of the investment in a relatively short time.
- False 14. Government competition in the construction of low-cost housing would probably not affect the value of slum property.
- True 17. In the past, housing has been one of the largest private industries in the United States.
- True 20. More than 50 per cent of the families in the United States have an annual income of \$1,800 or less; while at the same time over three-fourths of the houses built in the last five years were built to be sold for over \$4,000.
- False 22. Statistical studies show that cost of living is as high in suburban areas as in the metropolitan districts.

Reactions to these statements are summarized in terms of the number of accurate, inaccurate, and uncertain responses. These scores are used primarily for aiding the interpretation of scores on the first two sections of the test.

EVALUATION OF SOCIAL ATTITUDES

ANALYSIS OF THE OBJECTIVE

The study of social attitudes has been of concern to American psychologists and sociologists for a long time. The literature on this subject, however, reveals a great diversity of opinion regarding the proper delimitation of the behaviors to be called "attitudes" and the terminology to be used in denoting that behavior. Similar diversities also prevail in the conceptions of the important characteristics of "attitudes"

and in the techniques employed in measuring these characteristics.

The difficulties with the definition and classification of attitudes soon became apparent as the schools began appraising social attitudes. While the development of social attitudes was one of the most widely emphasized objectives among the schools in the Eight-Year Study, there seemed to be little clarity regarding the kind of behavior this objective involved and the significant areas in which it was important to develop and appraise social attitudes.

Analysis of Behavior

The initial statements from the schools revealed that many diverse types of behavior were considered to be social attitudes. Thus, some mathematics teachers submitted the ability to see quantitative relationships as an illustration of an attitude. Willingness to make an effort to express oneself clearly was one of the attitudes suggested by English teachers. Often objectives which seemed more closely related to interests and appreciations were included in this classification. Such personal qualities as resourcefulness, initiative in school work, and open-mindedness about the ideas of other people, along with beliefs about a wide range of social issues, were suggested in the statements of objectives submitted by the schools.

Recognizing the difficulties arising from the lack of clarity as to what kinds of behavior could be classified as attitudes and the diversity of objects toward which the suggested attitudes were directed, the committee on social attitudes proceeded along two major lines of analysis. It attempted (1) to describe the nature of social attitudes sufficiently to distinguish them from other school objectives, such as interests and appreciations, and (2) to delineate the major areas of social issues toward which social attitudes developed in school were usually directed. In doing this the committee

recognized that it could not solve the problem of defining and classifying attitudes in a comprehensive fashion. Since the committee was concerned with evaluation, it tried to identify only those aspects of social attitudes which constituted important objectives of the schools.

From this viewpoint the following distinguishing characteristics of attitudes were identified:

1. An attitude may involve a feelingtone of acceptance or rejection. This feelingtone may be evoked by an idea, a person, a way of behaving, or a mode of doing things. Thus one may like or dislike a person; reject or accept authoritarian methods; be afraid of or feel at home with members of the other sex, strange manners, or novel experiences. Attitudes of this sort are rather directly expressed in immediate behavior and the possession of "an attitude" may not necessarily be consciously recognized by the person concerned.

2. To have a belief about, or an opinion about, or to take position toward an issue, value, or institution may be considered another type of attitude. Thus one may approve of equality for Negroes, be for or against religion, disapprove of government control, believe in the efficacy of democratic processes, or be opposed to war. Though beliefs of this sort are not always arrived at by rational processes, they usually involve a conscious intellectual recognition that a position is being taken.

3. Often attitudes represent a latent tendency to act, such as the disposition to be kindly and considerate toward aliens, to defend the rights of minorities, or to proceed democratically in managing student government. Presumably these tendencies prevail as a result of conscious beliefs. However, this does not mean that there is of necessity a consistent relationship between what one believes and the character of overt action. Overt behavior may often be inconsistent with one's conscious beliefs, or it may express or imply value positions not consciously recognized as such by the individual.

Thus one may express prejudices toward certain ideas and values in one's daily behavior without reflecting upon the implications of these actions or without recognizing the beliefs which may have motivated them.

The problem of distinguishing the ways in which attitudes and social beliefs could be expressed was of major importance for purposes of evaluation, since these distinctions would largely determine the techniques to be used in appraisal. For this reason the relationship between "beliefs about" or "feeling toward" and overt action was discussed at length by the committee. Considered from the standpoint of the techniques to be used in appraisal of attitudes, the lists of specific attitudes submitted by the teachers suggested three groupings. Some of these objectives referred to attitudes pertaining to immediate social relations, such as cooperation and respect for others. The schools were concerned with attitudes of this sort primarily as expressed in some form of overt action. This type of attitude could therefore be appraised best by means of anecdotal recordings, behavior records, and observational checklists to be devised by each school for its own use.¹³

Another series of attitudes also permits expression in overt behavior, but social conventions and personal inhibitions tend to suppress that expression. Attitudes toward the other sex, toward family relations, toward certain aspects of one's own personality, and so on, are of this sort. Indirect methods of appraising these attitudes are necessary. A method of this type is described in the chapter on Personal and Social Adjustment.

¹³ Several such devices were developed. Behavior records developed under the leadership of Eugene R. Smith will be discussed in Part II of this book. The Francis Parker School developed a checklist, "Record for Describing Attitudes and Behavior in High School" covering: I, Cooperation; II, Responsibility; and III, Attitude toward School Work. A somewhat similar scheme for collecting anecdotal records was adopted in the Tower Hill School.

A third group dealt with such social issues as international relations, unemployment, freedom of speech, and democracy in school. While measurable consequences in overt behavior attend some of these attitudes, their expression is largely confined to a theoretical or verbal level. Even adults as individuals have only limited opportunities for expressing their beliefs through overt action. Thus, for example, belief in the desirability of government aid to agriculture would in the case of most people be expressed in verbal arguments, in taking sides on ideas presented in print, or in writing about these issues. Only such "token overt action" as writing to one's Senator or casting a vote on certain measures affecting the issue seemed to be open to the majority of people on a great many social issues. On the other hand, in a democracy the beliefs held by people influence social action by groups, and consequently a great deal of effort is directed toward clarifying beliefs and opinions on controversial issues. It was therefore thought important to appraise the development of these beliefs even though the appraisal would have to be confined to verbal expression of beliefs. Scales of beliefs inviting reactions toward statements of opinion on significant social issues seemed the most economical and appropriate method for appraising attitudes of this sort.

Areas of Social Beliefs

One of the first tasks in developing an instrument to evaluate social beliefs was to secure suggestions regarding the major areas of social beliefs to be covered in the appraisal. Obviously, it is possible to have a belief about almost anything, and almost anything can be covered by the term "social." It was clear also that certain of the possible areas of social beliefs were of more concern to schools than were others. The schools were, therefore, asked to suggest the areas of social beliefs in which they were interested. In several cases both students and parents as well as teachers par-

ticipated in this exploration. The rating scales and attitude tests already in use in schools were also examined. Samples of student writing were analyzed, as were their choices of "research" topics and free reading. In some classes daily logs of topics of discussion were kept.

When compiled, these suggestions included the following areas of social issues: democracy—political and economic, the role of the machine and invention in contemporary civilization, consumer problems, use of natural resources, labor, unemployment, housing, nationalism and internationalism, war and peace, school life, religion, and family. Some of these were mentioned by all schools and others by only a few.

In order to provide means of appraisal of so varied a range of social beliefs, a series of instruments was developed. With the exception of one instrument devoted to appraisal of beliefs on issues of school life, all of them deal with large social issues. The following list indicates the scope of this project.

1. Beliefs on Social Issues (Form 4.21-4.31), an instrument covering general social issues. Two forms were developed, one for the senior high school level, another for the junior high school.¹⁴
2. Beliefs on School Life (Form 4.6), an instrument covering issues in the area of school relationships.

These two instruments included issues which were suggested by a large number of schools and were designed for general use. In addition, several instruments were developed for more specific purposes. These included:

3. Beliefs on Economic Issues. This was made for a school particularly interested in developing economic attitudes through the study of selected short stories and poems.

¹⁴ Another form (4.9-4.10) included religion and family life in addition to the areas covered in Form 4.21-4.31.

4. A series of instruments sampling in detail beliefs on such issues as Men and Machines, Distribution of Wealth, Consumer Problems, and Use of National Resources, designed for a school emphasizing these particular problems.
5. Beliefs on Housing in your Community, for two schools conducting an intensive study of housing.

Of these, the development of the instrument Beliefs on Social Issues is discussed in detail in this chapter. Brief accounts are given of the Beliefs on School Life and Economic Beliefs.

EVALUATION OF BELIEFS ON SOCIAL ISSUES

Before an instrument suitable for appraising beliefs on social issues could be developed, it was necessary to (1) select the areas of issues to include, (2) determine the types of sub-issues to sample in each area, (3) decide on the level of intensity at which each of the statements in the test should be formulated, (4) designate the characteristics of beliefs which were to be measured, and (5) choose a technique appropriate for securing and summarizing the responses of students. This section summarizes the preliminary investigations which influenced the final decision on these problems.

Sampling of Issues and Formulation of Statements

From the list submitted by the teachers, six areas of interest to many schools were chosen by the committee. These were: democracy, economic relations, labor and unemployment, race, nationalism, and militarism. The problem of determining the specific issues to be sampled in each area and their specific direction was a more difficult one. To have a discriminating instrument, it is not only necessary to sample the significant aspects of an issue but also to sample the major variations in beliefs about these aspects. Each one of

the major areas chosen was broad enough to involve a host of more specific aspects. Thus the issue of equality of races involves such specifics as equality of work opportunity, of education, of political and civic rights, of social relations, and so on. A quite different set of sub-issues appears when the causes and consequences of racial equality or inequality are considered. The positions taken toward each of these aspects of racial equality may differ considerably in the case of the same individual, as well as from individual to individual. Thus those who believe that Negroes should have educational opportunities equal to those of whites may not believe that both groups should have equal opportunities for every kind of work.

For an effective appraisal of beliefs it is also important to determine a reasonable threshold for each statement. A statement of a position toward any social issue can be phrased with any degree of intensity. It can be phrased so strongly that very few people can agree with it, or so mildly that most people responding to it can agree with it. Thus, a statement expressing opposition to equality for Negroes could be phrased to deny any form of equality or permit only certain kinds of equality. A statement implying low standards of morality or lack of intellectual ability could be applied to all Negroes, or only to Negroes of certain social status, and so on. Effective statements for the purpose of the measuring instrument are ones which elicit a reasonable amount of both agreement and disagreement from the students.

Interwoven with this problem of threshold is the question of the use of language in the statement of beliefs. Because of the general nature of the issues, a certain degree of abstractness in stating them seemed unavoidable. Abstract terms, however, are often subject to different interpretations by different people. Statements of opinion frequently necessitate the use of emotionally colored words, the interpretation

of which varies from person to person. Care was therefore necessary to avoid words likely to be ambiguous to the students or likely to create emotional reactions causing an interpretation irrelevant to the intended meaning of the statement.

To get suggestions on how to deal with these problems, students in several schools were asked to submit statements of opinion on issues in each of the six areas chosen. Several hundred statements of opinion were collected in this way. A selection of these chosen from each area was resubmitted to the students. They were asked to indicate their agreement or disagreement with each of the statements and then to arrange all of the statements in ten groups, ranging from the ones they thought stated strong opposition to ones stating strong approval of the central issue in each area.

The results from these studies were used in several ways. By *a priori* analysis, lists of important issues to be sampled in each area had been drawn up. These lists were checked against the items suggested by the students to eliminate any issues of which students did not seem to be aware. The reduced lists of issues then served as a basis for formulating statements for the test. In the area of democracy, for example, the statements sampled the following issues:

1. Civil liberties, such as freedom of speech, the right to trial by a jury, and the right to vote.
2. Equality of opportunity and responsibility in a democracy, such as equality in economic and educational opportunities, and equality of responsibility in carrying the financial burden of the government.
3. Manner of appointing and electing government officials and representatives.
4. Functions and responsibilities of democratic government in promoting general welfare, such as providing medical care and social security for all.
5. Freedoms and responsibilities of citizens in a democracy.
6. Influences of social and economic classes on democracy.

From the students' responses it was also possible to determine the kinds of statements which were so extreme as to elicit either a unanimous agreement or a unanimous disagreement. Usually only the items on which there was a reasonable division of opinion were chosen. In a few instances, however, items were retained because they were considered important and because there was reason to believe that unanimity of opinion was caused by some special factor in the background of these students rather than by the fact that the issue was not in general a debatable or a significant one. Whenever possible, the terms used by students were employed. All statements were scrutinized by a jury of 12 persons for possible ambiguity, or other verbal difficulties, and for their relevance to the major issue.

Characteristics to Be Diagnosed

In considering the characteristics of beliefs, three were found to be of importance to schools. In the first place, the teachers wanted to see whether increased understanding of social problems brought about an ability and willingness to take personal positions upon an increasing range of social issues. One of the main criticisms of social education in schools had been centered on the failure to develop in students personal viewpoints toward important social issues. It was therefore decided that the prospective instrument should be so set up as to diagnose the extent to which students are able and willing to take a definite stand on social issues.

Teachers were also interested in learning about the *direction* of positions taken by the students. Thus they wanted to know whether on the whole students accepted or rejected the principle of universal freedom of speech, whether students were for or against certain measures to alleviate poverty and unemployment, and so on. This interest in the type

of positions taken did not imply a decision regarding the desirability of any *one* specific position, however. While there was a fairly close agreement among the teachers on the desirability of developing acceptance of democratic processes and of racial tolerance, it seemed both impossible and undesirable to classify the positions on many other issues as desirable or undesirable. At the same time, it seemed necessary to adopt some scheme of distinguishing and classifying the positions taken toward the statements of opinion included in the test. Unfortunately, most of the terms used to refer to the direction of attitudes suggest an idea of rightness or wrongness, approval or condemnation of a given position. The members of the committee wished to avoid such terms for summarizing the test results, but found it impossible to locate any terms which did not have such connotations. The terms *liberal* and *conservative* were finally adopted as a convenient way of describing two opposite directions on issues selected for the test. The meanings adopted for these terms will be discussed later in connection with the description of the scoring and summarizing of the responses.

The *consistency* of students' beliefs was a third characteristic teachers wished to diagnose. Teachers regarded consistency as a desirable characteristic of social beliefs, no matter which position was taken. The committee recognized at least two levels of consistency. Generalizing a multitude of specific beliefs in different areas into a coherent and consistent viewpoint represented one level. Inconsistency in this case would reveal itself by a shift of viewpoint from area to area. The other and more specific level involves the consistency of beliefs toward the same issue. Inconsistency in this case means agreement with expressions of opposite viewpoints on the same issue. It seemed possible to diagnose consistency of the first type by examining the direction of be-

liefs in each of the areas. To get evidence on consistency of the second type, two statements expressing opposite viewpoints on each issue were included in the instrument.

Techniques of Constructing the Scale

There are several possible techniques of securing and summarizing the responses of students to statements of issues. Thurstone regards the intensity of a feeling or position as the most significant characteristic of attitudes, and has developed a series of scales measuring the intensity of the favorable and unfavorable positions toward a single issue, such as war and peace.¹⁵ Each statement in a scale containing 20 or more represents a position toward a given issue, these positions ranging from intense opposition to intense approval, with a neutral zone in the middle. A quantitative "scale value" is assigned to each statement and the student's score is expressed as the median of the scale values of the statements he endorses. Low scores indicate opposition and high scores indicate approval. Another approach is used by Neumann.¹⁶ He attempts to combine a survey of various international issues with a measure of the intensity of reaction toward each one. He accomplishes this by including statements on a series of issues and by directing students to mark each statement by indicating five degrees of reaction ranging from strong approval to strong disapproval.

Although several schools in the Study used Thurstone's scale for measuring Attitude Toward War, and tried out experimentally a modified form of Neumann's Attitude Indicator, the committee decided that a still different technique would be more useful in serving the purposes of these

¹⁵ L. L. Thurstone and E. J. Chase, *The Measurement of Attitude* (Chicago, University of Chicago Press, 1929), pp. 10-12.

¹⁶ George B. Neumann, *A Study of International Attitudes of High School Students* (New York, Teachers College, Bureau of Publications, Columbia University, 1926).

particular schools. It was believed that separate scales, each of which focuses on a single major issue (e.g., war or religion), make it relatively easy for a student to decide what is likely to be the "acceptable" position and to respond accordingly, thus raising questions as to the validity of the instrument as an indicator of the student's "real" attitude. This aspect of validity might be at least partially protected by mixing statements on a variety of issues in the same instrument and avoiding the use of titles which would reveal the major issues included. Moreover, it seemed more important to the schools to appraise the positions on a *range of sub-issues* under each major area of issues than to scale in detail the *intensity* of each position. To attempt to do both would probably result in an instrument too long for practical use. All of these considerations influenced the technique which was eventually chosen and which will be described in the next section.

DESCRIPTION OF THE TEST ON BELIEFS ON SOCIAL ISSUES (FORM 4.21-4.31)

After the above-mentioned problems had been considered, a plan emerged for a new instrument to measure Beliefs on Social Issues. In the present form it consists of 200 statements, classified under the following areas of issues: democracy, economic relations, labor and unemployment, race, nationalism, and militarism. Students respond to each statement by indicating agreement, disagreement, or uncertainty. The statements are arranged in random order and are presented to the students in two sections given at different times. For each statement in the first section there is a statement in the second section representing the opposite point of view.

A sample of the statements is given below. The items from the two sections of the test are shown together. The key is inserted after each statement.

Democracy

- 4.21 1. Complete freedom of speech should be given to all groups and all individuals regardless of how radical their political views are.
(A, Liberal; D, Conservative; U, Uncertain.)
- 4.31 111. Freedom of speech should be denied all those groups and individuals that are working against democratic forms of government.
(D, Liberal; A, Conservative; U, Uncertain.)

Economic Relations

- 4.21 20. Since the welfare of a whole nation depends on its natural resources, their use should be subject to public control.
(A, Liberal; D, Conservative; U, Uncertain.)
- 4.31 125. Those who own oil wells, coal mines, and other natural resources should be allowed to operate them as they think best.
(D, Liberal; A, Conservative; U, Uncertain.)

Labor and Unemployment

- 4.21 14. Most workers who are unable to provide for themselves during a period of unemployment have been too shiftless to save.
(D, Liberal; A, Conservative; U, Uncertain.)
- 4.31 104. The wages of most workers are so low that it is impossible for them to save enough money to support themselves during periods of unemployment.
(A, Liberal; D, Conservative; U, Uncertain.)

Race

- 4.21 97. It is all right for Negroes to be paid lower wages than whites for similar kinds of work.
(D, Liberal; A, Conservative; U, Uncertain.)
- 4.31 192. The same wages should be paid to Negroes as to whites for work which requires the same ability and training.
(A, Liberal; D, Conservative; U, Uncertain.)

Nationalism

- 4.21 } 79. Our government ought to protect American business interests in foreign countries even if it involves using our army and navy.
(D, Liberal; A, Conservative; U, Uncertain.)
- 4.31 } 189. Our government should not risk a war to protect American business interests in foreign countries.
(A, Liberal; D, Conservative; U, Uncertain.)

Militarism

- 4.21 } 35. The amount of profit made from the sale of war materials should be strictly limited.
(A, Liberal; D, Conservative; U, Uncertain.)
- 4.31 } 132. Men should be allowed to make profits out of munition making just as they are allowed to make profits from other business enterprises.
(D, Liberal; A, Conservative; U, Uncertain.)

Scoring and Summarizing the Results

The responses to the whole test as well as to each of the areas are summarized under four main headings: liberalism, conservatism, uncertainty, and consistency. No attempt was made to arrive at a categorical definition of the terms *liberal* or *conservative*. These terms were adopted for convenience only and carry a somewhat different connotation with reference to each area. The *liberal* point of view in the area of democracy, for instance, tends to endorse freedom of speech; democratic processes in government; responsibility of the government for promoting the welfare of all groups in society with respect to health, security for old age, and the protection of consumers; and reinterpretation of the Constitution and other basic laws in keeping with present-day social and economic demands. The *conservative* position tends to approve restrictions on freedom of speech, to limit the responsibility of the government for social welfare, and to favor a strict interpretation of the Constitution.

In the area of economic relations, the *liberal* position tends

to endorse government regulation of public utilities, natural resources, wage levels, insurance, and to approve of moving in the direction of production for use rather than for profit. The *conservative* position represents the policy of economic individualism, the policy of *laissez faire*, and the preservation of the profit system in unrestricted form.

With respect to labor and unemployment, the *liberal* position tends to favor collective bargaining; to approve of social legislation providing for minimum wage levels, health insurance, and unemployment relief; and to maintain that unemployment is caused by social conditions beyond the control of individuals, and hence that its consequences should be borne by society rather than by the individuals who happen to be affected by it. The *conservative* position tends to oppose the organization of labor for collective bargaining; to oppose labor legislation or expenditure of government funds for relief of unemployment; and to maintain that unemployment is caused by some deficiency of the individuals, and hence that the consequences should be borne by those who happen to be unemployed.

In the area of race, the *liberal* position tends to endorse the equality of all races as far as social, economic, and educational opportunities are concerned, and to deny that racial inequality is inherent or inborn. The social, economic, and educational status of Negroes as a group is attributed to environmental conditions rather than to hereditary causes. The *conservative* position accepts the inherent supremacy of the white race and indorses racial discrimination of all sorts.

A pacifistic viewpoint represents *liberalism* in the area of Militarism: that is, the tendency to favor arms limitation, arbitration, and condemnation of war as a way of settling international troubles. Belief in the inevitability of war, in armed preparedness, in the use of armed force, and in the benefits of military training for character development illustrates the *conservative* position.

In the area of nationalism, a *liberal* viewpoint is ascribed to those who are internationally-minded, who recognize the worth and the contributions of other nations, and who deny that there is need for protecting a nation's imperialistic economic enterprises abroad with armed forces. A *conservative* viewpoint is associated with emphasis on national glory and honor, and the belief that American ways would be best for other peoples; it tends to defend the notion of the supremacy of America and of things American in all respects and to insist on the use of American standards in judging the contributions of other nations.

In all areas the *uncertain* response is taken to mean either that the student does not understand the statement or that he is unable to take a position regarding the issue because of conflicting ideas about it. It was also anticipated that a relatively high degree of uncertainty might characterize the position of the more thoughtful students. *Consistency* indicates the extent to which students take a similar position twice on the same issue; i.e., do not agree with both of two contradictory statements. The tendency to take a similar position on a range of issues in one area or in different areas is indicated by the percentage of *liberal* and *conservative* responses in each area.

As can be seen from the data sheet, these four headings (liberalism, conservatism, uncertainty, and consistency) are used to summarize both the total scores and the subscores for each of the six areas. No such headings are used in the instrument itself, and the student is not aware that his responses are to be classified in this way. Moreover, it cannot be emphasized too strongly that, as far as the instrument is concerned, there is no implication that either the *liberal* or the *conservative* position is to be preferred. The instrument is designed to measure the status or change of beliefs; the problem of determining the desirability of the *direction* that

the beliefs of students should take is a responsibility of the schools.

Explanation of the Data Sheet

The scores on this scale can be interpreted in terms of three questions centering about the direction, uncertainty, and consistency of the viewpoints shown. The first question is: *What is the direction of the pattern of beliefs and how is it distributed?*

The scores on *liberalism* (columns 25 and 1-6) indicate the per cent of the statements to which the student responded in the liberal direction.¹⁷ The scores on *conservatism* (columns 26 and 7-12) give the per cent of responses made in the direction described as conservative. High scores in either direction, uniformly distributed, would mean a fairly well-thought-out position. Student A, for example, has responded in the liberal direction to 90 per cent of all items in the test (column 25). His scores, furthermore, are distributed evenly in all of the six areas (columns 1-6). Student D achieves a similarly high and fairly even distribution of scores in the conservative direction (columns 7-12).¹⁸ Student B is near the median of the class as far as his total score on liberalism is concerned, but there is a good deal of fluctuation of his liberal responses from area to area. He is, for instance, inclined to an international viewpoint (N, 80) and pacifism (M, 78), but is at the same time inclined to reject collective bargaining and social measures to combat unemployment (LU, liberalism, 44, conservatism, 50). The same type of fluctuation can be observed in the scores on liberalism of Student C. In the area of economic relations (ER, 7)

¹⁷ The terms "liberal" and "conservative" are used throughout this section in place of making the more lengthy references to their meaning in each of the areas.

¹⁸ The usual ratio of liberal to conservative scores in the schools of the Eight-Year Study was about 2:1. Hence scores on conservatism which are as large as, or larger than, scores on liberalism were interpreted as a high degree of conservative beliefs.

SAMPLE DATA SHEET

School Experimental

(Four illustrative cases from a class of 33)

Summary for Test 4.21-4.31

Grade 12

"Scale of Beliefs"

Column Number →	% Liberalism						% Conservatism						% Uncertainty						% Consistency						Totals %*																							
	D			ER			LU			R			N			M			D			ER			LU			R			N			M			Lib			Con			Unc			Cons		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28																				
Student: A	88	96	94	92	89	83	8	4	3	8	9	8	4	0	3	0	2	9	84	96	84	80	76	75	90	6	4	84																				
B	70	53	44	58	80	78	24	37	50	16	14	22	6	10	6	26	6	0	60	44	40	80	60	65	63	28	9	57																				
C	44	7	14	88	53	67	14	33	30	12	6	2	42	60	56	0	41	31	64	32	60	75	44	60	43	16	41	56																				
D	34	24	39	32	16	15	65	76	61	58	83	80	1	0	0	10	1	5	56	80	72	40	56	70	28	70	2	67																				
Maximum Possible	100 per cent in all columns																																															
Lowest Score	34	7	14	12	16	15	0	0	0	0	3	0	0	0	0	0	0	1	0	32	24	40	15	24	25	28	2	0	32																			
Highest Score	96	100	97	100	89	93	65	76	61	58	83	80	46	66	56	50	54	47	88	100	92	100	84	85	93	38	43	90																				
Median	70	53	64	88	53	65	14	17	11	8	20	12	16	20	22	8	22	20	60	48	64	75	52	60	61	16	22	57																				

*These are scores showing response to the test as a whole, and are not sums of the sub-scores distributed by areas.

and labor and unemployment (LU, 14) he makes few liberal responses, while his score on toleration of racial equality (R, 88) is very high. In his case, however, the absence of liberal responses cannot be interpreted as a rejection of this position. His scores on uncertainty in these two areas (uncertainty: ER, 60, LU, 58) indicate that in these areas he has difficulty in taking a position. In the few instances he does so, the responses in the conservative direction prevail (ER: liberal, 7, conservative, 33; LU: liberalism, 14, conservatism, 30).

The second question is: *To what extent are the students willing (or able) to take definite positions on these social issues?*

The *uncertainty* (columns 27 and 13-18) scores give the per cent of responses in which a student neither agrees nor disagrees with the statements. High uncertainty might mean desirable caution, inability to understand the statements, lack of information, or lack of conviction. In most cases this response seems to mean "I don't know or I can't decide," for socially conscious and active students usually have low "uncertain" scores. Thus, Student C is very uncertain of his position on all of the issues with the exception of those pertaining to race. He scores far above the median for the class on total uncertainty (column 27), and in five of the areas. Students A and D indicate little uncertainty as to their positions. Neither extreme certainty nor extreme uncertainty in themselves are desirable. Whether or not either can be considered desirable depends on the total pattern of scores. Thus, certainty combined with high consistency is more acceptable than high certainty combined with low consistency because flexibility is important as long as there is confusion. Experience with test data has shown also that certainty combined with high conservatism is not as desirable from the standpoint of growth as is high certainty combined with high liberalism. This conclusion was drawn because it was

found that conservative beliefs were more frequently borrowed beliefs, while liberal beliefs were more often arrived at through personal thought and consideration. In interpreting the meaning of high or low uncertainty, however, the developmental trend of the student needs to be considered. Thus one would expect an increase in uncertainty whenever an individual is in a state of transition from one type of social viewpoint to another.

The third question is: *To what extent are the students consistent in the positions they take?*

The *consistency* (columns 28 and 19-24) scores give the per cent of consistent responses on the total test and in the areas listed above. High scores in these columns indicate clarity of outlook, whether it is liberal or conservative in its direction. Low consistency may occur for at least two reasons. Students may be inconsistent because of inability to think through their beliefs or because they are actually embracing conflicting positions. In the first case, there is likely to be an even distribution of inconsistency scores in all areas. In the other case there is more likely to be high consistency in some areas and low consistency in other areas. While high consistency can be generally regarded as a desirable characteristic, one must be aware that often inconsistency is a by-product of transition from one pattern of beliefs to another. In the latter case, low consistency may be an index of change and may be temporary. Whether this is true or not can be determined if the test is readministered after an appropriate interval of time and a description of the kinds of changes taking place in students is secured.

Student A is the most consistent of the four students whose scores are given on the data sheet. Student B shows a variable pattern of consistency. On racial issues he is rather consistent (consistency: R, 80), but on issues of labor and unemployment he is the least consistent student in his entire group (consistency: U, 40). Similar fluctuation in consistency from

area to area is shown by Students C and D. Student D is rather consistent on issues in economic relations (consistency: ER, 80) and relatively inconsistent on racial issues (consistency: R, 40).

The scores on liberalism, conservatism, and uncertainty are interdependent and must be viewed in relation to each other. This can be illustrated by comparing the scores on economic relations for Students C and D. Both of these students have low scores on liberalism in this area, but while Student C is rather highly uncertain, Student D is highly conservative. Thus scores on liberalism alone tell only part of the story. One can infer that the low score on liberalism in the case of Student C results from the fact that he has not made up his mind on many of the issues. Student D, however, seems to have definite convictions about economic relationships. For this reason the interpreter must, in addition to studying each score independently, consider the whole pattern of scores before arriving at a final judgment about a student or groups of individuals.

Several other general considerations apply in interpreting different combinations of score patterns. Thus, when the score on uncertainty is unusually high, the scores on both liberalism and conservatism are of necessity low. In such cases one can interpret these scores better by comparing them with each other than by comparing each with the median. Thus, in the case of Student C one might say that whenever he makes up his mind on economic relations his position will be predominantly in the conservative direction, because 33 per cent of the items are marked in the conservative direction while only 7 per cent of the items are marked in the liberal direction. High scores on uncertainty, coupled with high scores on consistency, are more likely to be an indication of intelligent doubt than of mere confusion and inability to see the issues clearly. Conversely, lack of uncertainty where inconsistency is high would indicate a pre-

mature feeling of security about beliefs which in reality are confused. Decisions such as these concerning the relative desirability of high or low scores on liberalism are left for the teacher to make.

Although in the course of the above discussion comments have been made concerning the scores of four students, no attempt has here been made to present a complete and coherent account of the beliefs of these students. The data on each student and each group of students made available by this instrument are too extensive to permit the presentation within the limits of this chapter of a complete treatment of the possibilities of interpretation.

Validity and Reliability

Several factors influence the validity of this instrument. In the first place, there is the problem of the role of language in expressing feelings and viewpoints. In statements of issues terms which have different meanings for different individuals are apt to be used. The expressions of attitudinal positions also require the use of some words or ideas to which strong emotional reactions are attached and these reactions usually are not the same from individual to individual. Certain words may evoke responses somewhat independent of, or irrelevant to, the meaning and intent of the whole statement. Also involved is the fact that many individuals are not clear about their own beliefs. Those who tend to be confused or uncertain about their own positions are apt to respond more or less automatically to familiar terminology in place of attempting to decide what their own beliefs are. Moreover, it is likely that individuals with no definite beliefs on a given issue may be induced to give definite responses merely because familiar verbal stereotypes are presented to them.

Secondly, there is the problem of securing honesty of response. Social beliefs are somewhat in the realm of the pri-

vate life of an individual and he is not always willing to reveal them. There are either general social pressures or pressures in a given group toward the "right way of believing," and individuals whose personal beliefs differ from the predominant ones may feel threatened in disclosing them. Thus, often in a school where the majority of students are *liberal* in a certain respect, those who do not share the liberal viewpoint are put on the defensive. This applies also to teacher-pupil relations. Even in responding to an instrument of this sort which is not, properly speaking, a "test," students are apt to try to live up to the expectations of teachers who are known to favor certain viewpoints rather than to express their own viewpoints. It is for reasons like these that the question of validity is peculiarly complex in the measurement of social beliefs.

An additional difficulty lies in the fact that the social beliefs of individuals are rarely so generalized that the subjects mentioned in the statements do not affect the response. Thus, in securing opinions upon the issue of government control vs. economic individualism, it may make a considerable difference whether the issue is stated with reference to public utilities or to railroads, whether the object of control is profits or wages, and so on. Ideally, the specific issues used in the test should include *all* of these variations. Since this ideal cannot be achieved in a test of this sort, one is faced with the problem of sampling and of the reliability of the sample.

The efforts made in the process of construction to assure high validity for the test were described above. Summarized briefly, these consisted of securing a clear delimitation by the committee of the behavior to be measured, and of utilizing statements from students in deciding which specific issues to include, in determining the level of intensity at which statements should be formulated, and in phrasing the statements. Finally, the instrument was revised several times

on the basis of analyses of student responses to tentative forms.

In addition to the above precautions, several studies of the validity were conducted.¹⁰ In the first study the instrument was given to 65 junior and senior classes studying American history and sociology in a large public high school. Verbal descriptions of the beliefs of these students based on their numerical scores were made and these were discussed with the cooperating teacher. The validity of the scores in each area in the scale was considered separately. The teacher's judgments of the social attitudes of the students as revealed by his observations in the classroom coincided with the interpretations of the scores from the test in 90 per cent of the cases.

Thirty of these students were interviewed. They were chosen on the basis of the test scores so that they represented the ten most conservative, the ten most liberal, and the ten most inconsistent and uncertain students in the entire group. The questions asked in these interviews paralleled the statements of the test. Some of the students were questioned regarding their points of view within a single area; others were interrogated with respect to two, three, or even all six areas. When the information obtained in this way was compared with the test results, the two sets of data were found to be fairly consistent; that is, the direction of points of view, the certainty, and the consistency of the students as revealed by the test were very closely related to those indicated by their verbally expressed opinions.

A second study of the validity of the instrument was carried out in a ninth grade social science class composed of 18

¹⁰ These validity studies were conducted by Paul R. Grim and the discussion here summarizes his findings described in more detail in "A Technique for the Evaluation of Attitudes in the Social Studies," a dissertation submitted to the Ohio State University in 1939. Dr. Grim's study was made in connection with the Form 4.2-4.3. Only slight revisions were made in the Form 4.21-4.31.

students. Written descriptions of their social beliefs as revealed by test scores were made. Apprentice teachers collected hundreds of anecdotes pertaining to expressions of, and behavior relative to, the social viewpoints of the students, and also examined these students' written work. They then summarized their findings by rating these students on a five-point scale for liberalism and for consistency in each of the six areas. It was found that over 90 per cent of the judgments of the teachers coincided with the test ratings. The students in this group were also interviewed. In 17 out of the 18 cases, the opinions expressed in the conferences conformed closely with the responses to the test.

In one study of reliability, coefficients²⁰ for this test based on a total population of 600 students selected from 14 schools and representing grades nine through twelve were computed. The results were as follows: On liberalism they ranged from .79 to .86 for the different areas; for the total score on liberalism the coefficient was .95. On conservatism they ranged from .72 to .81 in different areas; the reliability coefficient for the total score on conservatism was .93. On uncertainty the range of reliability coefficients was from .79 to .85, and a coefficient of .96 was obtained for the total score. On consistency the reliability coefficients ranged from .45 to .61, with a coefficient of .85 for the total score.²¹ These data check rather closely with those obtained in other studies from other populations and by other methods. The scores in the test are stable enough so that, within appropriate statistical limits, they may be used for diagnosis of individual as well as group differences.

As can be seen from these data, the stability of the scores by areas is a good deal lower than the stability of the total scores. The scores on consistency by areas have particularly

²⁰ Estimated by the Kuder-Richardson formula. More complete data on reliability and other statistics are given in the Appendix.

²¹ Since pairs of items are scored to determine consistency, the test is in effect only half as long for this purpose.

low stability and can be used only to designate the extremes. All other scores used within the content of the whole pattern of scores and within appropriate statistical limits, can be used for helpful diagnostic judgments regarding the nature of social beliefs.

BELIEFS ABOUT SCHOOL LIFE

Another scale of social beliefs (Beliefs about School Life, Form 4.6), was devoted to the area of school life.

Appraisal of the beliefs regarding various aspects of school life was considered important for several reasons. In the first place, students' points of view on such matters as grades and awards, methods of teaching, and ways of conducting the school government, determine to a considerable extent the type and the effectiveness of their adjustment to school. The beliefs prevailing among students on these matters also influence the organization and functioning of the school since students' beliefs play an important part in motivating their behavior in specific situations. Finally, certain of these beliefs represent aspects of "democracy in school" and as such are considered in many schools as desirable ends in themselves. Awareness of the nature of these beliefs on the part of both students and teachers is helpful in accomplishing desirable changes in the school environment or in an individual student's reactions to that environment. For these reasons a means of obtaining systematic evidence on beliefs toward a range of issues about school life was thought to be a desirable addition to observations of overt behavior.

Analysis of the Objective

In order to be sure that the test sampled opinions on issues of concern relative to school life, two investigations were conducted. First, some students were asked to write brief essays on "Democracy in My School." Their essays discussed many kinds of problems, from rules regarding the use of lip-

stick to criticism of the course of study which they were following. Secondly, a list of the major areas of school life and illustrative statements of issues in each area was sent to teachers in several schools. They were asked to criticize the choice of issues and the tentative list of specific statements, and to make additions to either if they thought there were important omissions. In analyzing the material obtained from teachers and students, it was found that the most frequently mentioned issues could be classified in six major areas: school government, curriculum, grades and awards, school spirit, pupil-teacher relations, and group life. These became categories of summary for the instrument which was developed. This instrument is similar in form to the one described in the preceding section except for the difference in content and the fact that no attempts were made to measure consistency. It consists of a series of 118 statements of opinion, and students respond by indicating either agreement or disagreement with them, or uncertainty about them. In the following paragraphs a brief description of the categories and some illustrative statements from the instrument are given.

Description of the Test

The area of *school government* samples such issues as appropriate bases for electing students to school offices, treatment of minority groups, appropriate degree of student responsibility for the conduct of school affairs. Student responses to these items are classified as democratic and undemocratic. For example, agreement with each of the following statements is scored as a "democratic" response, and disagreement with these statements is scored as an "undemocratic" response:

19. Criticisms of the school government made by first year pupils should be considered just as carefully as criticisms which juniors and seniors make.

20. The teachers and principal should have pupils help in deciding what books to buy for the school library.

The area of *group life* involves issues of the status of various school groups and their relations to each other and to school. The following problems are included: the extension of special privileges of various sorts only to members of certain groups, the maintenance of class distinctions in terms of these groups, and the desirability of characterizing students as members of certain groups or cliques rather than as individuals. Responses to these items are summarized in terms of the number of responses indicating a "social attitude," meaning approval of equal treatment of all groups, and a "class" attitude, indicating a disposition to approve all kinds of distinctions and cliques. For example, agreement with the following statements indicates a "class" attitude, whereas disagreement indicates a "social" attitude:

6. Pupils from the wealthier families in a community and pupils from the poorer families should not be put in the same homeroom together.
99. In most cases, it is undesirable to have slow and bright pupils working together in the same class.

The area of *pupil-teacher relations* involves problems of sharing responsibility between teachers and pupils, and of the methods by which the allocation of responsibility should be made. The following issues are sampled: the appropriate degree of pupil-planning of various school activities, methods of making decisions, types of problems which teachers alone should solve. Reactions to this group of items are summarized in terms of the number of responses indicating approval of cooperative relations, and the number indicating approval of authoritarian relations. Following are two illustrations of items in this area in which disagreement with the item indicates approval of cooperative methods and agreement indicates approval of authoritarian methods:

2. It is better for a teacher to decide what the pupils are to study in a class than to let the pupils plan their work by themselves.
17. Too much time is wasted when pupils take part in the discussion of plans for a unit of study.

The area of *curriculum* involves issues of educational philosophy and practice. Responses to these issues are summarized in terms of liberal and conventional attitudes. A "liberal" attitude is indicated by an experimental point of view: that is, a belief in the integration of school subjects, pupil-teacher planning, flexibility in planning units of study, and in utilizing community resources. A "conventional" attitude is indicated by a disposition to maintain rigid subject matter divisions, to prefer teacher-planned courses of study, and to emphasize the acquisition of facts and information. The following statements are taken from this area:

11. It would be a good idea for several teachers of different school subjects to take part in a class discussion with a group of pupils.
56. Trips outside of the school building should not be taken at a time when they interfere with the regular class schedule.

In the above illustration, agreement with the first statement indicates a "liberal" attitude, whereas agreement with the second indicates a "conventional" attitude toward school problems.

The area of *grades and awards* samples issues concerning the appropriate use of grades and awards, and the types of grades and awards which are desirable. For example, such statements as the following are made:

18. If a pupil receives failing grades most of the time, it shows that he is not learning anything in school.
50. If grades were done away with, pupils would have no

way of knowing whether they were making progress in their studies.

Responses to such issues are summarized as non-traditional or traditional. "Non-traditional" attitudes are indicated by questioning the desirability of using grades and awards as incentives, as means of determining participation in school activities, and as providing the exclusive measure of the value derived from school life. The "traditional" point of view is indicated by an acceptance of grades and awards for such purposes.

The area of *school spirit* is sampled by issues concerning the extent of school loyalty which is desirable, and the types of expressions of school loyalty which are appropriate. For example, the following statements are offered for consideration:

40. We would get some helpful ideas for improving our school by visiting other schools to see how they do things.
102. One of the best ways for a pupil to show that he is a good school citizen is always to defend his school when others criticize it.

Agreement with the first statement is classified as a "cosmopolitan" point of view, agreement with the second as a "provincial" attitude. A "cosmopolitan" viewpoint is indicated by a disposition to recognize certain weaknesses in one's own school, a disposition to view the school as a changing rather than as an inflexible institution, and a tendency toward "worldliness" in one's relations with students from other schools. A "provincial" viewpoint is indicated by expressing intense loyalty to one's immediate group to the extent of excluding cooperative relations with other groups. In addition to the descriptive categories, the number of uncertain responses in each area is given.

As is indicated by the method of summarizing student responses, the test may be useful in identifying points of view

on the part of an individual student which are likely to be hampering his adjustment to, and active participation in, school life. It must be noted, however, that the test has not been studied sufficiently to warrant a recommendation that it be used for precise individual diagnosis. Its primary usefulness is for studying groups. Only students who deviate markedly from the group pattern can be identified with assurance as being significantly different from others in the group.

A teacher who wishes to use the test should examine it with respect to her own school situation in terms of the following criteria: (1) Does it sample problems and conflicts which pupils in this school must deal with in order to make a better adjustment to school life? (2) Are the beliefs toward school life which are sampled likely to affect participation in social movements and processes outside school? (3) Does it involve issues regarding educational philosophy which are really controversial issues within this school? (4) Does it sample beliefs which may provide clues concerning the behavior of individual pupils in a variety of situations in this school?

BELIEFS ON ECONOMIC ISSUES

Frequently the Evaluation Staff received requests for specialized instruments to evaluate certain unique features of a particular school program. One such request was for the development of means of appraising the effects on social awareness of the reading of fiction dealing with social problems. The literature used in this program described social and economic problems, offered explanations of the causes and effects of these conditions, and suggested (in certain cases) types of solutions for the problems.

Analysis of the Objective

In analyzing the effects of such a program, it was apparent that they might be classified as follows: (1) increasing

student *awareness* of existing social and economic conditions; (2) stimulating the development of a consistent *social philosophy*; and (3) aiding students to see the *implications of their personal social philosophy for concrete action in specific problem situations*.

Two characteristics were thought important in describing *awareness* or recognition of social and economic conditions. First, there is the extent of the awareness or lack of it. The extent of awareness may be characterized either by the range of problems of which an individual is aware or by the depth of understanding about any particular problem. It was decided that in this instance the range of problems to which an individual responds was more significant than the depth of his understanding of any one problem. The lack of awareness may be expressed in several ways. Students may believe that conditions are worse than facts indicate, that they are better than the facts indicate, or they may feel uncertain about either the existence or non-existence of these conditions. The second characteristic of awareness is consistency. An individual who has a clear impression of actual social and economic conditions will not agree with both of two plausible statements describing exactly opposite conditions. An instrument designed to measure awareness of social and economic conditions should yield evidence on each of these characteristics of awareness.

An individual's *social philosophy* may also be described in terms of several characteristics. First, there is the question of its general direction: Is it highly individualistic? Is it based on humanitarian values and considerations of general welfare? Is it dominated by the acceptance of the *status quo*? Does it indicate a willingness to change contemporary social and economic conditions? Second, the degree of certainty with which an individual holds a particular point of view is of interest in appraising his social philosophy. Certainty may be defined either with respect to one's degree of conviction about any single issue or with respect to the range of issues

toward which one indicates a positive point of view. For the purposes of this particular appraisal, certainty in the latter sense was considered more significant. The third important characteristic of a social philosophy is the degree of its internal consistency.

An individual's *ability to see the implications of his social philosophy* for concrete social action may be described first with respect to the predominant type of social action he generally approves or disapproves in specific problem situations, and in terms of the variety or comprehensiveness of things which he agrees should be done. Second, the type of social action about which he is frequently uncertain can be described. Third, the types of problem situations in which he approves an extensive and far-reaching social action, those in which he approves little or no social action, and those in which he is primarily uncertain, may be indicated.

Description of the Test

On the basis of the analysis of (a) the types of issues sampled in the literature and of (b) the nature and characteristics of the behavior to be measured, a test called Scale of Beliefs on Economic Issues was constructed. This test is made up of three parts.

The *first part* of the test consists of statements that certain conditions do or do not exist in the United States. The statements are made in pairs so that while one statement indicates the existence of a given condition, the other statement in the pair indicates the existence of exactly the opposite condition. The student reacts by indicating that he agrees, disagrees, or is uncertain about each statement purporting to describe existing conditions. In order to get an index of the consistency of his responses the two scales containing opposite statements are given on different days. Responses to this part of the test are summarized in terms of the number of answers which indicate awareness of social

and economic conditions, lack of awareness of these conditions, uncertainty about them, and consistency of belief about them.

The *second part* of the test consists of statements sampling various points of view regarding the types of conditions which are desirable. These statements are also made in pairs in order to obtain evidence on the consistency of the student's social philosophy. One set of conditions, if considered desirable, implies approval of the *status quo*; whereas the other, if followed to its logical implications would involve changes in the present scheme of things. The issues sampled in this section of the test parallel those sampled previously. That is, in the first section there is a statement as to the extent to which people achieve economic security today, in the second section, a statement concerning the degree to which people ought to have economic security. The student reacts by indicating agreement, disagreement, or uncertainty about each statement. A student's responses to this section of the test are summarized in terms of the degree to which he accepts and approves the *status quo*, the degree to which he accepts a social philosophy which implies change in the present order, the degree to which he is uncertain about his social philosophy, and the degree to which his social philosophy is internally consistent.

The *third part* of the test is made up of a number of problem situations describing some specific instances of the conditions described in the first section of the test. The description of the problem is followed by five courses of action that represent different points of view about what should be done about such specific problems. The types of points of view sampled in the courses of action have been labeled *futile*, *conservative*, *compromise*, *liberal*, and *radical*. These terms are not to be understood as meaning anything other than convenient summaries of various points of a scale ranging from the attitude of "do nothing" to the attitude of "change the

whole system." The student is asked to indicate whether he agrees, disagrees, or is uncertain about each course of action. His responses are summarized in such a way as to indicate the extent to which he agrees, disagrees, or is uncertain about each type of social action.

USES OF THESE TESTS

The fact that a test is valid "in general" does not assure that valid results are necessarily obtained in a given school or with a given group of students. There are many conditions which must be fulfilled if these tests are to be useful. The most obvious one is that the teacher should be interested in developing the kinds of behavior diagnosed in the test. Thus the tests dealing with social values and beliefs should be considered only if the development of social beliefs and the ability to analyze social problems in terms of a personal pattern of social values is of concern to the school.

A certain minimum background on the part of the students is also assumed in several of these tests. For instance, to obtain valid results from the test on Social Problems (Form 1.42), it is necessary for students to have had some opportunity to discuss controversial problems, to develop viewpoints with reference to them, and to acquire familiarity with basic democratic values. Otherwise their responses will be conditioned by factors other than their ability to apply value principles, such as lack of familiarity with these principles. Similarly the test Application of Social Facts and Generalizations (Form 1.5) is explicitly designed for use with students who have had opportunity to study issues similar to the ones used in the test and have acquired some general information about them. Occasionally a teacher may want to use an exercise from this test as a pre-test, before undertaking a specific unit of study. This is appropriate when the students have had some general experience with the problem

and the teacher is anxious to find out at which level to attack the problem with them.

It is also important for the teacher to decide whether the content and vocabulary of these tests are appropriate for his group.²² Too often in selecting a test, consideration is given only to its appropriateness for a given grade level. Pupils who do not respond sensitively to the connotations of the words used in these tests will not give an accurate picture of their social beliefs and values. The absence of a time limit helps, but not sufficiently for many groups.

The attitudes and expectations of students at the time of taking the test regarding the purpose of the test and the use of the results are extremely important in all tests in which students are expected to express their own viewpoints. If the students expect to be graded on such tests, or if for some reason they think that they should please the teacher, they are likely to mark the test according to their best guess of what is expected of them. Certain precautions have been taken in the tests themselves to prevent dishonest marking. Thus in the Scales of Beliefs the items pertaining to a range of issues are in random order to make it more difficult for the students to see what the "acceptable" responses might be. In the Social Problems test the directions for marking the test do not reveal the kind of analysis to be made of the responses. No such precautions, however, can take the place of a classroom in which the pupils and the teacher trust one another.

Provided, then, that the qualities diagnosed in the test are of concern to the teacher, that the content and vocabulary of the tests are appropriate to the level of student development, and that students feel free to express their own views, several fruitful uses of the results are possible. In the first

²² With the exception of the test, Beliefs on School Life, which can be used in grades seven to twelve, none of these tests is appropriate for non-verbal students, nor should they be given below the tenth grade except in unusual cases.

place, the teacher may want to *diagnose the strengths and weaknesses of the individuals in his class*, in order that he may give each one the kind of help he needs. In the case of the application of social values, the difficulty of some students may be in their lack of social awareness, while others are blocked by their inability to see the implications of social values in concrete social problems. Conflicting or confused values prevent clear thinking for some students, while gullibility to slogans may be the main difficulty with others. Each needs a different kind of help. Experiences necessary for broadening awareness do not necessarily contribute to greater consistency. The methods employed to clarify values and beliefs and to eliminate prejudices differ from the methods of building up a more realistic understanding of social phenomena. Students whose difficulty is the absence of any personal viewpoint are not helped by the kinds of experiences needed by those handicapped with entrenched biases and prejudices. The results of the test on Social Problems (Form 1.41 or 1.42) throw some light on the needs of individuals in these respects.

If the teacher is interested in the development of social beliefs, he may want to know in which areas students tend to be confused, to embrace conflicting viewpoints, or have unfounded prejudices. Information of this type may also serve as a background for understanding difficulties in thinking logically. For example, students who reveal strong prejudices in the area of economic relations in the Scale of Beliefs test often make mistakes in reasoning in this area in the Social Problems test. The barrier is emotional, not necessarily intellectual.

Through the use of these tests, the teacher can also *check the effectiveness of his curriculum*. For example, the study of current social problems was introduced in many schools in the hope of engendering social awareness and a greater ability and inclination to use scientific methods in dealing

with social phenomena. First-hand exploration of the community and use of literary materials to illustrate social problems became a part of most programs. Democratic processes in administering school affairs were introduced in the hope that personal democratic attitudes might be developed. These hypotheses need to be checked by evidence of changes taking place in students. Furthermore, curriculum experiences effective in one respect sometimes produce unexpected and undesirable results in some other respect. Thus, courses dealing with modern problems, introduced to enlarge social awareness, sometimes increase inconsistency and enhance ambivalence and confusion of social values. An emphasis on democratic processes in school may develop loyalty to certain values in this situation, but without proper reference to larger social problems, a double standard of democratic values may result.

There are many points at which an objective check is particularly needed. One of the most common difficulties in social education is that students tend to master generalized concepts without seeing concretely enough how these concepts apply in a variety of life problems. Thus, students tend to remember and accept such democratic tenets as equality of opportunity or freedom of speech, without recognizing in life the problems in which these values are involved and the ways in which they are violated. The use of the Scale of Beliefs in conjunction with the test on Social Problems shows in what degree these difficulties are present among students.

A teacher may also want to see whether his students are achieving an increasingly *consistent social viewpoint*. Most individuals tend to accept values which are in conflict with others which they hold at the same time. While one would not expect anyone to be wholly free of these conflicts, one would hope that with increasing maturity and with increasing understanding these conflicts would tend to be eliminated. Often, however, school programs tend to increase

these conflicts rather than to eliminate them. This is particularly the case when the community or the family has a different philosophy from the one emphasized in the school.

A similar effect is produced when students are exposed to many new experiences creating new beliefs and values without sufficient time to reconsider the values they have already developed in their previous experiences. Conflicts are particularly apt to appear between general beliefs and their specific implications. Thus, it is not uncommon to see students approve of a more equitable distribution of wealth in general and at the same time be violently opposed to such practical measures to achieve it as the graduated income tax or minimum wage law. As long as the school programs tend to emphasize generalities, while experiences at home and in the community contribute to the development of specific values and loyalties, such conflicting viewpoints are unavoidable. An increasing ambivalence and conflict rather than increasing clarification and integration of social outlook result unless teachers are continually aware of points at which individuals need help in integrating or clarifying their value concepts and beliefs. The examination of the distribution of the scores on values in the social problems test and of the scores on scales of belief would reveal to what degree and at which points individuals and groups are embracing contradictory values and beliefs.

In addition to diagnosing the strengths and weaknesses of individuals at a given time, teachers may also be interested in changes occurring over a period of time. The diagnosis of growth is particularly important in connection with the aspects of social sensitivity dealt with in this chapter. Changes in fundamental value patterns, methods of applying values, and using information to gain deeper insight into complex social problems do not take place overnight. The results of experiences at a given time may not show up until a good deal later. Moreover, these are objectives which cannot be

finally established during the high school years. At best, one can hope to establish certain tendencies and predispositions and to initiate certain techniques of analysis and inquiry. This means that it is important to get evidence of the direction of changes taking place in students. Administering tests of this sort over a period of time would help determine such long-term changes.²³

Generally it is not advisable to use any of these tests less than a year apart. They are too general in content, in the first place, to reveal minor changes. Secondly, the scores are not reliable enough to detect small amounts of change. However, the exercises in the test Application of Social Facts and Generalizations (Form 1.5) can be used as a pre-test and as an end-test in evaluating the effectiveness of a given unit of study, within an interval of a few weeks. The use of these exercises as a pre-test would serve two ends: (1) to diagnose the background of the students in order to attack the problem at an appropriate level, and (2) to give direction and impetus to the study. The end-test would show how well students had mastered the ideas and techniques for understanding a given problem.

It must be pointed out here that while each of these tests was designed as an independent unit, better information about the students and the effectiveness of the curriculum is secured when several of them are given and interpreted together. This is particularly true of the Scale of Beliefs (Form 4.21-4.31) and of the Social Problems test (Forms 1.41 and 1.42). These two tests were planned as companion instruments—one to give an overview of general beliefs, and the other to diagnose their application in concrete situations. In most cases the data from a single instrument must

²³ The tests of beliefs, such as the Scale of Beliefs on Social Issues (Form 4.21-4.31) can be administered several times. Two forms of the test on Social Problems (Forms 1.41 and 1.42) have been made available. These forms are sufficiently similar to enable teachers to compare scores on one form with those on the other form.

be supplemented with other evidence before safe inferences can be drawn. This is particularly the case when it is necessary to carry the diagnosis to the point of locating the causes of difficulty. Thus, ambivalence of value pattern may be the result of lack of acquaintance with the issues involved, lack of ability to see logical relations, sheer inability to read and to understand this test, or a genuine division of viewpoint. These possibilities have to be checked against other evidence, such as reading scores, scores on psychological tests, tests on logical thinking, or daily observations of students' behavior in the classroom. Only after such checking can the teacher be safe in planning the experiences necessary to eliminate the difficulties.

In still other cases, the interpreter needs to resort to a more detailed analysis of student responses than is possible by examining the score sheet. In the case of the Social Problems test, some students may have difficulties in connection with certain problems and issues and not with others. Whenever there is reason to believe that the scores on the data sheet have covered up important information, it is profitable to examine the answer sheets themselves.

Chapter IV

ASPECTS OF APPRECIATION



INTRODUCTION

All of the lists of objectives submitted by schools in the Eight-Year Study mentioned the development of a wide range, an increasing depth, and a personal selection of interests and appreciations. Accordingly, an interschool Committee on the Evaluation of Interests and Appreciations was formed early in the Study and met frequently to analyze this area of objectives. One of its first conclusions was that, although interests and appreciations are so closely related that it is often impossible to distinguish them in specific instances, techniques for evaluating them would be sufficiently different to justify a division of labor. The committee was therefore divided into sub-groups after arriving at a common understanding of the objectives to be considered. Many subtle distinctions were drawn between interests and appreciations, but their common purport seemed to be that interests emphasize "liking" an activity, while appreciations include "liking" but emphasize "insight" into the activity: understanding it, realizing its true values, distinguishing the better from the worse, and the like. The sub-committees on appreciations developed instruments chiefly in the fields of literature and the arts, which are reported in this chapter. The work of the Committee on Interests is reported in Chapter V.

APPRECIATION OF LITERATURE

Since there are somewhat different points of view as to what is meant by the objective "Appreciation of Literature,"¹ it is important to recognize at the outset that the analysis which will be described here is restricted to an analysis of certain types of students' reactions to reading. This restriction should not be taken to imply that other behaviors might not be included under the heading "Appreciation of Literature"; a number of articles and studies might be cited to illustrate the range of behaviors which have, at various times, been identified with appreciation. Carroll,² for example, mentions information, sensitivity to style, understanding of "deeper meanings," and emotional response as included in appreciation. In developing his tests of prose appreciation Carroll chose to measure students' ability to differentiate the good from the less good and the less good from the very bad.³ This ability has been regarded by many as an important element in, or index of, appreciation. Logasa and Wright, to cite a second example, have made a rather extensive analysis of appreciation⁴ and have published tests of the following behaviors: discovery of theme, reader participation, reaction to sensory images, discrimination between good and poor comparisons, recognition of rhythm, and appreciation of fresh expressions as opposed to triteness. Instead, the restriction mentioned above merely implies a selection, on the part of the committee, of behaviors which (1) were regarded by them as important aspects of appreciation, and (2) were not being adequately appraised by the available instruments. A major question which the committee

¹ Cf. Broom, M. E., "Literature and Aesthetics," *The High School Teacher*, VIII (October, 1932), pp. 293-294.

² Carroll, Herbert, "A Method of Measuring Prose Appreciation," *English Journal*, XXII (March, 1933), p. 184.

³ *Op. cit.*, p. 185.

⁴ See "Tests for Measuring Appreciation," *School Review*, XXXIII (September, 1925), pp. 491-492.

wished to be able to answer is: "How do students react to their reading?" For convenience, certain of these reactions to reading have been designated as "Aspects of Appreciation."

The Committee's Analysis of Students' Reactions to Reading

The Committee on the Evaluation of Reading was organized in the fall of 1935. In selecting members for this committee the schools recognized that teachers other than teachers of literature are often responsible for guiding the reading of students and hence should participate in the evaluation of reading outcomes. For this reason, in addition to the field of English, other areas, such as social studies, the core program, the school library, and school administration, were represented by various members of the committee. Because of the wide geographical distribution of the schools in the Eight-Year Study, this committee was divided into two sub-committees, one of which met in New York City and the other in Chicago. During the school years 1935-36, 1936-37, and 1937-38 a number of committee meetings were held in these two cities. The meetings held in New York City were attended by representatives of 16 eastern schools; meetings in Chicago were attended by representatives of eight schools in the Middle West. Members of the Evaluation Staff also attended these meetings and coordinated the work of the two sub-committees.

The Committee on the Evaluation of Reading undertook, as its first task in developing instruments for appraising students' reactions to their reading, to clarify what was meant by "reactions to reading." A preliminary analysis of students' reactions to reading was made, at the request of the committee, by Carleton Jones of the Evaluation Staff and was submitted to them for revision. After some discussion, the committee selected from the preliminary analysis seven behaviors

or reactions to reading which seemed to them to be of considerable importance. These are:

1. *Satisfaction in the thing appreciated*

Appreciation manifests itself in a feeling, on the part of the individual, of keen satisfaction in and enthusiasm for the thing appreciated. The person who appreciates a given piece of literature finds in it an immediate, persistent, and easily-renewable enjoyment of extraordinary intensity.

2. *Desire for more of the thing appreciated*

Appreciation manifests itself in an active desire on the part of the individual for more of the thing appreciated. The person who appreciates a given piece of literature is desirous of prolonging, extending, supplementing, renewing his first favorable response toward it.

3. *Desire to know more about the thing appreciated*

Appreciation manifests itself in an active desire on the part of the individual to know more about the thing appreciated. The person who appreciates a given piece of literature is desirous of understanding as fully as possible the significant meanings which it aims to express and of knowing something about its genesis, its history, its locale, its sociological background, its author, etc.

4. *Desire to express one's self creatively*

Appreciation manifests itself in an active desire on the part of an individual to go beyond the thing appreciated: to give creative expression to ideas and feelings of his own which the thing appreciated has chiefly engendered. The person who appreciates a given piece of literature is desirous of doing for himself, either in the same or in a different medium, something of what the author has done in the medium of literature.

5. *Identification of one's self with the thing appreciated*

Appreciation manifests itself in the individual's active identification of himself with the thing appreciated. The person who appreciates a given piece of literature responds to it very much as if he were actually participating in the life situations which it represents.

6. *Desire to clarify one's own thinking with regard to the life problems raised by the thing appreciated*

Appreciation manifests itself in an active desire on the part of the individual to clarify his own thinking with regard to specific life problems raised by the thing appreciated. The person who appreciates a given piece of literature is stimulated by it to re-think his own point of view toward certain of the life problems with which it deals and perhaps subsequently to modify his own practical behavior in meeting those problems.

7. *Desire to evaluate the thing appreciated*

Appreciation manifests itself in a conscious effort on the part of the individual to evaluate the thing appreciated in terms of such standards of merit as he himself, at the moment, tends to subscribe to. The person who appreciates a given piece of literature is desirous of discovering and describing for himself the particular values which it seems to hold for him.

An example may aid in clarifying each of these seven behaviors. Let us suppose that a student has read a particular novel, such as Dickens' *Tale of Two Cities*, and that during the reading of this book he has read attentively and with absorption (1). Let us also suppose that he has derived such satisfaction from the book that he plans to read it again and to read other novels by Dickens (2). Perhaps his curiosity about Dickens as an author, about the literary currents of the middle nineteenth century, about the historical novel as a type, or about the French Revolution has been aroused by his reading (3). He might want to sketch Carton riding to the guillotine or try to conceive in words some scene or character which grows out of his reading (4). While reading he might "lose himself" in the events of the book, he might, like Booth Tarkington's Willie Baxter, become one with Carton and feel that "It is a far, far better thing that I do . . ." (5). Many problems might be suggested or raised again for him by his reading; he might want to think through what

friendship or love implies, what the proper ends of life are, what terror and force effect in the world (6). Finally, he might want to compare this novel with others by Dickens and others of its type, compare his judgments of it with those of other persons, seek out its values and its limitations (7).

This statement of important reactions to reading is a selective one and should be regarded as such. A number of other reactions or responses to reading might be identified and judged to be of importance by other teachers or test makers. Pooley,⁵ for example, has made a rather detailed analysis of "fundamental" and "secondary" responses to prose and poetry which differs somewhat from the analysis accepted by the committee. Since our purpose is to report what was done by these committees and the Evaluation Staff during the period of the Eight-Year Study, a comprehensive discussion of the many definitions of appreciation or of the many possible analyses of responses to reading cannot be given. Consequently, the omission of a careful consideration of the many studies and tests of literary appreciation which have been made by others should not be regarded either as an oversight or as evidence of a belief that the work reported here exhausts the topic "The Evaluation of Appreciation of Literature."

Instruments Which Were Developed to Appraise Students' Reactions to Their Reading

A number of instruments were developed for the evaluation of students' reactions to their reading. Three of these instruments make use of a questionnaire technique which consists essentially of asking students to observe themselves, in retrospect, and to record these observations. This technique was arrived at in the following manner. The committee first discussed ways in which the seven types of reaction

⁵ Pooley, Robert, "Measuring the Appreciation of Literature," *English Journal* (High School Edition), XXIV (October, 1935), pp. 627-633.

to reading might be manifested in readily observable student behavior and prepared a list of overt acts and verbal responses which, they judged, would in certain situations reveal the presence or absence of each of these seven types of behavior. A few of the overt acts and verbal responses which were included in this list are:

1. *Satisfaction in the thing appreciated*
 - 1.1 He reads aloud to others, or simply to himself, passages which he finds unusually interesting.
 - 1.2 He reads straight through without stopping, or with a minimum of interruption.
 - 1.3 He reads for considerable periods of time.
2. *Desire for more of the thing appreciated*
 - 2.1 He asks other people to recommend reading which is more or less similar to the thing appreciated.
 - 2.2 He commences this reading of similar things as soon after reading the first as possible.
 - 2.3 He reads subsequently several books, plays, or poems by the same author.
3. *Desire to know more about the thing appreciated*
 - 3.1 He asks other people for information or sources of information about what he has read.
 - 3.2 He reads supplementary materials, such as biography, history, criticism, etc.
 - 3.3 He attends literary meetings devoted to reviews, criticisms, discussions, etc.
4. *Desire to express one's self creatively*
 - 4.1 He produces, or at least undertakes to produce, "a creative product more or less after the manner of the thing appreciated.
 - 4.2 He writes critical appreciations.
 - 4.3 He illustrates what he has read in some one of the graphic, spatial, musical, or dramatic arts.
5. *Identification of one's self with the thing appreciated*
 - 5.1 He accepts, at least while he is reading, the persons, places, situations, events, etc., as real.

- 5.2 He dramatizes, formally or informally, various passages.
- 5.3 He imitates, consciously and unconsciously, the speech and actions of various characters in the story.
- 6. *Desire to clarify one's own thinking with regard to the life problems raised by the thing appreciated*
 - 6.1 He attempts to state, either orally or in writing, his own ideas, feelings, or information concerning the life problems with which his reading deals.
 - 6.2 He examines other sources for more information about these problems.
 - 6.3 He reads other works dealing with similar problems.
- 7. *Desire to evaluate the thing appreciated*
 - 7.1 He points out, both orally and in writing, the elements which in his opinion make it good literature.
 - 7.2 He explains how certain unacceptable elements (if any) could be improved.
 - 7.3 He consults published criticisms.

The committee next suggested that one method of securing evidence of these seven types of response in secondary schools would be to ask students to report on these behaviors themselves. The advantage of asking students to observe themselves and to record these observations, as compared with the collection of anecdotal records or the use of interviews, is primarily one of practicability. The committee also recognized that the use of a questionnaire technique demands that certain assumptions be fulfilled if the method is to give valid evidence. Most important among these assumptions are: (1) that the overt behaviors and their accompanying situations specified in the items are significant evidence of the seven types of behavior; (2) that the students are capable of observing these overt behaviors, of remembering them, and of recording them; (3) that the students are honest in their responses to each item. The extent to which these assumptions actually are fulfilled will depend upon both the characteristics of the questionnaire itself and the

situation in which the student is asked to respond to the questionnaire. First, let us review the construction of one of these three questionnaires, pointing out the criteria in its construction which were made necessary by these assumptions; later we shall consider the administration of such an instrument and the conditions under which its use is most apt to give valid evidence.

Questionnaire on Voluntary Reading

Of the three appreciation questionnaires—The Novel Questionnaire, The Drama Questionnaire, and The Questionnaire on Voluntary Reading—which were developed during the period of the Eight-Year Study, The Questionnaire on Voluntary Reading was used and studied most extensively; for this reason it will be chosen to illustrate the construction of an instrument to measure students' responses to their reading. This questionnaire was designed to measure the extent to which students exhibit the seven types of response to their "free" or voluntary reading of books. The directions to the student on the questionnaire read in part as follows:

QUESTIONNAIRE ON VOLUNTARY READING

Directions to the Student

The purpose of this questionnaire is to discover what you really think about the reading which you do in your leisure time. Altogether there are one hundred questions. Consider each question carefully and answer it as honestly and as frankly as you possibly can. *There are no "right" answers as such.* It is not expected that your own thoughts or feelings or activities relating to books should be like those of anyone else.

The numbers on your Answer Sheet correspond to the numbers of the questions on the questionnaire. There are three ways to mark the Answer Sheet:

- A—means that your answer to the question is *Yes*.
- U—means that your answer to the question is *Uncertain*.
- D—means that your answer to the question is *No*.

If it is at all possible, answer the questions by *Yes* or *No*. You should mark a question *Uncertain* only if you are unable to answer either *Yes* or *No*.

Please answer *every* question

One hundred questions which the student is asked to answer make up the items of the questionnaire. An illustrative set of items, grouped under the seven types of response,⁶ follows:

"Derives satisfaction from reading"

1. Is it unusual for you, of your own accord, to spend a whole afternoon or evening reading a book?
2. Do you ever read plays, apart from school requirements?

"Wants to read more"

1. Do you have in mind one or two books which you would like to read sometime soon?
2. Do you wish that you had more time to devote to reading?

"Identifies himself with his reading"

1. Have you ever tried to become in some respects like a character whom you have read about and admired?
2. Is it very unusual for you to become sad or depressed over the fate of a character?

"Becomes curious about his reading"

1. Do you read the book review sections of magazines or newspapers fairly regularly?
2. Do you ever read, apart from school requirements, books or articles about English or American literature?

"Expresses himself creatively"

1. Have you ever wanted to act out 'a scene from a book which you have read?
2. Has your reading of books ever stimulated you to attempt any original writing of your own?

"Evaluates his reading"

1. Do you ordinarily read a book without giving much thought to the quality of its style?

⁶ In the questionnaire itself, the items are ungrouped; they are, however, readily classified by use of the scoring key.

2. Do you ever consult published criticisms of any of the books which you read?

"Relates his reading to life"

1. Has your attitude toward war or patriotism been changed by books which you have read?
2. Is it very unusual for you to gain from your reading of books a better understanding of some of the problems which people face in their everyday living?

It will be observed that this statement of the seven types of behavior differs somewhat from that given on pages 251 and 252. The major purpose of this rewording was to place the emphasis, for several of these types of behavior, on what students actually do rather than on what they desire to do.

The first criterion that the items included in the questionnaire had to satisfy was that they must deal with behaviors which were judged by teachers who prepared and used the questionnaire to be significant evidence of the seven types of response to reading. In a sense, then, the items constitute a definition, in terms of what students do and say, of what these teachers meant by "Derives satisfaction from reading," "Wants to read more," etc. In order to insure that this criterion was satisfied, the items were drawn originally from the list of overt acts and verbal responses which the committee judged to be significant evidences of the seven types of response. Then, as use of the questionnaire in a number of schools gave opportunity to secure from teachers additional judgments of the significance of these items, the questions were revised.

In selecting and phrasing items it was necessary to consider several additional criteria. The assumption that students are capable of observing these overt behaviors in themselves, of remembering, and of recording them demands first of all that each item deal only with those behaviors which secondary school students are apt to exhibit and only with situations in which students are apt to find

themselves. This is almost an obvious criterion, for if we expect the student to report on his behavior we must ask him questions about things he actually has an opportunity to do. The committee, in preparing the list of overt acts and verbal responses, and teachers, in judging the significance of items included in the early forms of the questionnaire, were asked to consider whether or not each of the specific acts or verbal responses is something which secondary school students are apt to do or say. It was possible later, by studying the responses of students to each item on the questionnaire, to check these judgments of teachers to some extent. Second, this assumption demands that each item deal with behavior and situations which the student is apt to remember. This criterion immediately rules out certain types of questions. In general, we would not expect students to remember, for example, exactly how many books they had read during the summer; yet we might expect them to remember whether or not they had read a book during the preceding week. In general, we would not expect them to remember the details of an argument with a friend about the merits of a particular book; yet we might expect them to remember having tried to defend their judgment of a book. Third, this assumption demands that any judgments or generalizations which the student is asked to formulate be relatively simple ones. An item which calls for an extensive introspection, for the rating of one's self on an abstract and undefined quality, for making fine distinctions between causes or effects, etc., thus would be ruled out. Fourth, this assumption demands that each question be so phrased that it is readily understood by the student and can be answered with a minimum of writing. That the question must be understood if he is to answer it intelligently is obvious. That his ability to express himself in writing may become a factor which, for this test, may inappropriately condition the evidence and the judgments made from the evidence, was also recognized. The selection of

Yes, *No*, and *Uncertain* as the particular pattern of "controlled response" for the questionnaires eliminated the necessity of the student's writing out his answers, but made it necessary that each question be so phrased that it could be answered with one of the three responses provided.

The assumption that students are honest in their responses also suggests criteria which each item must meet. Certain activities and certain situations may have such a "prestige" value that questions dealing with them would tempt the student to say that he took part in them, whether he actually did or not. Questions dealing with any activity which is ordinarily participated in because of its "social" value thus were ruled out, as were all questions dealing with activities in which participation might be dependent primarily upon an economic factor. Likewise, items which deal with activities or situations, the disclosure of which might threaten the student's sense of security, may tempt him to disavow actual participation in these activities or situations. Questions which asked students to admit the reading of certain kinds of materials which are commonly frowned upon, such as comic magazines, or to disclose any of his more intimate feelings or relationships with other persons also were ruled out. The final criterion for the selection of the items, then, is that they deal only with overt acts and verbal responses which the student might be expected to report honestly.

Summarizing and Scoring the Questionnaire on Voluntary Reading

Several forms of the Questionnaire on Voluntary Reading were prepared during the period of the Eight-Year Study; comparison of these several forms reveals that (1) the items included in Form 3.32 probably best meet the criteria outlined above, (2) the length of Form 3.32 probably is an optimum for both practicability and reliability,⁷ (3) the

⁷ Statistical data on reliability are presented in the Appendix.

method of summarizing Form 3.32 is statistically preferable. For these reasons, the form of the Questionnaire on Voluntary Reading which is recommended for use is Form 3.32.

Form 3.32 is made up of the set of directions reprinted on page 253 and a list of 100 questions which students are asked to answer with one of three responses: *Yes*, *No*, or *Uncertain*. The responses to each of these 100 items are summarized under six categories: (1) Likes to read, (2) Identifies himself with reading, (3) Becomes curious about reading, (4) Expresses himself creatively, (5) Evaluates his reading, (6) Relates his reading to life. Originally, seven categories were used for summary of the scores on the questionnaire, but study of the students' responses revealed that scores on the categories "Derives satisfaction from reading" and "Wants to read more" are so closely related statistically as to warrant their being consolidated under one heading, "Likes to read." On page 259 there is presented a sample of the data sheet on which the scores made by individual students on Form 3.32 are reported. The scores of five students are presented for purposes of illustration. At the bottom of the data sheet appear the maximum possible score for each column, and the highest, the lowest, and the median score for each column computed for the class from which these five students were selected. All the scores on the data sheet are expressed as per cents; for example, the scores in column one are per cents of the 35 responses which are grouped under the heading "Likes to read."

Three scores are available for each of the categories: an "Appreciation" score, a "Non-appreciation" score, and an "Uncertain" score. For each category the "Appreciation" score summarizes the responses which indicate that the student engages in those behaviors which are regarded as significant evidence of that type of behavior; the "Non-appreciation" score summarizes the responses which indicate that the student does not engage in those behaviors; and the "Uncer-

School ExperimentalGrade 12

SAMPLE DATA SHEET

Summary for

Test 3.32

	Part I "Likes to Read"			Part IIA "Identifies"			Part IIB "Curious"			Part IIC "Expresses"			Part IID "Evaluates"			Part II Total			Part III "Relates to Life"			Total Score		
	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un	Ap	Non	Un
Column Number →	1	2	3	5	6	7	9	10	11	13	14	15	17	18	19	21	22	23	25	26	27	30	31	32
Student: A	98	2	0	80	0	20	80	0	20	70	20	10	40	20	40	68	10	22	64	4	32	77	6	17
B	66	20	14	40	30	30	60	30	10	50	40	10	90	10	0	60	28	12	64	32	4	63	26	11
C	88	12	0	80	0	20	90	0	10	60	30	10	90	10	0	80	10	10	80	20	0	83	13	4
D	42	32	26	40	10	50	0	80	20	0	30	70	10	40	50	13	40	47	16	32	52	24	35	41
E	58	42	0	50	50	0	10	90	0	20	80	0	20	80	0	25	75	0	36	64	0	39	61	0
Maximum Possible	100 per cent in all columns																							
Low Score	30	2	0	40	0	0	0	0	0	0	0	0	10	0	0	13	0	0	16	0	0	24	2	0
High Score	98	62	26	100	50	50	100	90	30	100	80	70	100	80	60	100	75	47	100	68	52	98	61	41
Median	72	26	4	80	10	30	50	40	10	50	40	20	80	20	10	63	30	4	60	32	4	64	27	4

tain" score gives the proportion of items which he was unable to answer with either *Yes* or *No*. In addition to these scores for each of the six categories, total "Appreciation," total "Non-appreciation," and total "Uncertain" scores may be computed. These total scores summarize the responses to all the 100 items of the questionnaire and are analogous to the "single score" given by many tests.

An explanation of the scores made by these five students follows:

Part I. *Likes to Read*

Columns 1, 2, 3 *Column 1* gives the per cent of responses which reveal that the student likes to read. *Column 2* gives the per cent of responses which reveal that he does not like to read. *Column 3* gives the per cent of uncertain responses. A high score in column 1, accompanied by low scores in columns 2 and 3, indicates that the student likes to read to a great extent. Student A, for example, has such a score. Low scores in columns 1 and 3, accompanied by a high score in column 2, indicate that the student dislikes reading. Among these five students, Student E has the highest score in column 2; however, reference to the line marked "High Score" reveals that his score in column 2 is not the highest in this class. A high score in column 3, such as that of Student D, indicates that the student was somewhat uncertain in answering the questions grouped under this heading.

Part IIA. *Identifies*

Columns 5, 6, 7 These scores indicate the extent to which the student identifies himself with his reading. Among these five students, Students A and C have relatively high "Appreciation" scores on this category (*column 5*) and zero "Non-appreciation" scores (*column 6*). Such scores indicate that the student identifies himself with his reading to a considerable extent. Student E has the

highest "Non-appreciation" score on this category, both among these students and among the class as a whole. Student D has a high "Uncertain" score (*column 7*).

Part IIB. *Curious*

Columns 9, 10, 11 These scores indicate the extent to which students are curious about their reading. Students A and C have high "Appreciation" scores (*column 9*) and low "Non-appreciation" scores (*column 10*). This pattern indicates that these students respond to their voluntary reading by wanting to know more about authors, books, literary periods, etc. Students D and E probably do not respond in this fashion, for they have low scores in column 9 and very high scores in column 10. *Column 11* gives the per cent of responses marked "Uncertain."

Part IIC. *Expresses*

Columns 13, 14, 15 These scores indicate the extent to which the student expresses himself creatively as a response to his reading. The highest "Appreciation" score (*column 13*) in this class is 100; none of these five students has such a high score in column 13; Students A and C are somewhat above the median of the class (50), and Student B is at the median. Probably none of these five students expresses himself creatively to a very great extent. Student E, with his high "Non-appreciation" score (*column 14*), probably rarely engages in such activities as creative writing, painting, dramatizing; etc. Student D is characterized by a very high "Uncertain" score (*column 15*).

Part IID. *Evaluates*

Columns 17, 18, 19 These scores indicate the extent to which the student evaluates or judges his reading. Students B and C have high "Appreciation" scores (*column 17*) and low "Non-appreciation" scores (*column 18*); this pattern indi-

cates that they tend to evaluate their reading to a very great extent. Student A has a low score in column 17, as compared with the median, and his "Uncertain" score (*column 19*) is rather high. This pattern differs considerably from the pattern of his scores on the preceding categories, and it suggests as an hypothesis that his greatest weakness may be a failure to engage in such activities as reading reviews and criticisms, attempting to make judgments about what he reads, etc.

Part II. *Total*

Columns 21, 22, 23 These three scores represent the totals of the scores in the four preceding categories and are reported primarily to provide measures whose reliabilities are comparable to those of the scores on Parts I and III. For the group of responses included in Part II, student C has a relatively high "Appreciation" score (*column 21*) and relatively low "Non-appreciation" (*column 22*) and "Uncertain" (*column 23*) scores. In diagnosing the specific differences between him and Student A, for example, it is necessary to refer to the four preceding categories. Student D has the lowest "Appreciation" score and the highest "Uncertain" score on Part II; Student E has the highest "Non-appreciation" score.

Part III. *Relates to Life*

Columns 25, 26, 27 These scores indicate the extent to which the student relates his reading to his life and to the problems which he recognizes as existing. A high "Appreciation" score (*column 25*), such as that of Student C, indicates that he relates his reading to life, as he knows it, to a considerable extent. Student E has a high "Non-appreciation" score (*column 26*), in fact almost the highest in the class. Probably he does not relate his reading to life to any great extent. Students A and D have rather high "Uncertain" scores (*column 27*).

Total Score

Columns 30, 31, 32 These scores are convenient for making a summarizing judgment of a student's responses to the test; however, they necessarily obscure some of the differences among students on various categories. The "Appreciation" score (*column 30*) gives the number of the student's responses to the one hundred items of the test which reveal these seven reactions to reading; the "Non-appreciation" score (*column 31*) gives the number of his responses which reveal that he does not react to reading in these seven ways, and the "Uncertain" score (*column 32*) gives the number of his uncertain responses.

Several rather commonly occurring patterns are revealed by the scores of these students. A set of scores which reveals that the student responds to his reading to a considerable extent in these seven ways is illustrated by that of Student C. Nearly all his "Appreciation" scores are relatively high and his "Non-appreciation" and "Uncertain" scores relatively low. Almost the opposite pattern is revealed by the scores of Student E: relatively low "Appreciation" scores and relatively high "Non-appreciation" scores. The relatively high "Uncertain" scores of Student D reveal that, despite the instructions to answer the questions with *Yes* or *No* if it were at all possible, he answered a large number of the questions with *Uncertain*. Several hypotheses might be advanced to account for this: He may have been quite indifferent to the test and have marked almost at random; he may have been extremely "overcautious" or scrupulous in attempting to answer the questions; he may have been unable to answer many of these questions because he had failed previously to observe such behaviors in himself. Further study of other data about this student would be necessary to confirm or deny these hypotheses and to arrive at a satisfactory interpretation of such

a pattern of scores. The scores of Student A indicate a student who likes to read very much yet does not evaluate his reading to any great extent. His relatively high "Uncertain" scores on Part IID and Part III should be used as a starting point for hypotheses as to why he responded in this fashion only to these two categories.

Other Instruments

Two questionnaires, similar in structure to the Questionnaire on Voluntary Reading, were developed for the purpose of measuring students' responses to a particular novel or a particular drama which they have read. The Novel Questionnaire (Test 3.22) includes 65 items, the responses to which are summarized under the same six categories as are the responses to Form 3.32. Similar scores are computed for each of the six categories, and for the total of 65 items. The Drama Questionnaire (Test 3.21) includes 80 questions, the responses to which are summarized under the six headings mentioned above plus an additional heading: "Feels that he understands the play." This category was added to the Drama Questionnaire in order to aid in the interpretation of scores on the six categories. It was believed that the extent to which a student feels that he understands the play he has read may demand differing interpretations of his other responses. For example, a pattern of scores which indicates that a student derived no satisfaction from reading the play yet felt that he understood it perfectly probably would demand a different interpretation from one which indicates that the student derived no satisfaction from reading the play and felt that he did not understand it. A similar category has not been added to the Novel Questionnaire; it is possible that teachers using the Novel Questionnaire would find such an addition helpful.

Each of the three questionnaires described includes, as has been indicated, a set of items the responses to which are summarized under the heading, "Evaluates his reading." The

purpose of this category is to discover to what extent students actually engage in such activities as comparing the merits of one book with those of another, discovering what critics have said about books they have read, comparing their judgments of books with those made by others, etc. Scores on this category obviously do not furnish information about the *quality* of the judgments which the student makes of books, just as scores on the category "Likes to read" do not furnish information about the *quality* of the books which he actually reads. Because a number of teachers wished to have some objective means of appraising the quality of students' judgments, this evaluation problem was explored. Three experimental instruments were developed; these are: An Interpretation of Literature (Test 3.1), Critical-Mindedness in the Reading of Fiction (Test 3.7), Judging the Effectiveness of Written Composition (Test 3.8). Because these instruments have not been used extensively or studied sufficiently, they are not as yet to be recommended for widespread use. However, they might serve as useful classroom exercises and they might suggest techniques for appraising students' judgments which others would want to utilize.

These three tests use short stories as their content or subject-matter. In brief, they were constructed by first asking a group of students to write out any judgments of the story which they could or would care to make. After these judgments had been sorted and the duplicating ones discarded, they were submitted to a jury of teachers. The jury grouped them and marked each as a "good" or a "poor" judgment. The test was then made up, including the story and the list of students' judgments, and those who took the test were directed to read the story and respond to each of the judgments listed by agreeing with it, disagreeing with it, or stating that they could neither agree nor disagree. The evaluation of each judgment made by the jury is used as a test key. Scores are given in terms of the extent to which the

student evaluated these judgments as did the jury. It should be pointed out that this is only one method of scoring responses on such a test. Other methods might be devised which would better suit the purposes of particular schools or teachers.

Test 3.1, *An Interpretation of Literature*, is based on O. Henry's story "A Municipal Report." The student is asked, after reading the story, to respond to statements which are grouped under such headings as:

1. What is your interpretation of the story?
2. What was O. Henry's point of view?
3. What was O. Henry's philosophy?
4. What was the character's motive?
5. Which is the most logical ending for the story?

Scores for each of these parts may be computed.

Test 3.7, *Critical-Mindedness in the Reading of Fiction*, makes use of two short-short stories reprinted from a popular magazine. The statements which follow each of these stories deal with the extent to which the actions and speech of these characters, the description given by the authors, the outcomes of the stories, etc., are "true to life." For example, these statements follow the story "First Acquaintance" by I. A. R. Wylie:

1. The general atmosphere—the smells, the signs on the door, the moving nurses, etc.—is depicted accurately in this story.
2. It seems scarcely likely that a young man would wonder about the "No visitors" sign, the oxygen tank, and the sick mother and daughter as the youth in this story did.
3. Under the circumstances it seems natural for the youth to say "Gosh" and "That's tough" several times.
4. No nurse, even a young one, would volunteer as much information about patients to a stranger as the nurse in this story does.
5. The youth's sudden realization of what death means and

- his thoughts about his own mother seem real and natural.
6. The suggestion that the youth was crying when he left the hospital is difficult to believe.
 7. The emphasis upon the fact that the mother and daughter were alone in the world seems exaggerated and over done.
 8. Under the circumstances it seems natural for the young man, on his return to the hospital the next morning, to be more concerned to find out about the condition of the sick girl's mother than of that of his sister.
 9. The action of the young man in going into the girl's room to tell her that she had not been left completely alone is in accordance with what the reader has previously found out about his character.
 10. The sick girl's response to his sympathy does not seem true to life.

Four scores are given on this test: (1) "Judicious," i.e., the extent to which the student's responses agree with the jury's judgment; (2) "Hypercritical," i.e., the extent to which the student judges situations which the jury believes are true to life to be not true to life; (3) "Uncritical," i.e., the extent to which the student judges situations which the jury believes are not true to life to be true to life; (4) "Uncertain," i.e., the extent to which the student was unable to agree or disagree with these statements.

Test 3.8, Judging the Effectiveness of Written Composition, makes use of a short-short story written by a high school student. This story is followed by 28 statements about the narrative quality, the style, the characterization, etc., of this story. For example, these statements are included:

1. The writer should not have included so many different episodes in one brief story.
3. The writer shows considerable skill in depicting the humorous aspects of situations.
4. The dialog in the story is, in general, handled ably.
5. Esmond's stammering, hesitant way of speaking in trying

situations helps the reader to see him as an individualized character.

6. The concluding episode provides a very effective climax for the story.
7. *Esmond* is a good name for the chief character in the story.

This test is also scored by comparing the student's responses with those provided by a jury of adults.

Validity of the Questionnaires

In order to assess the value of the instruments designed to measure students' responses to their reading it will be necessary to consider their validity, their reliability, and the uses which classroom teachers may make of them. It was pointed out earlier that the validity of the questionnaire technique for measuring students' responses to their reading is primarily dependent upon the extent to which three major assumptions are fulfilled; it was also pointed out that whether or not these assumptions are fulfilled will depend upon both the nature of the instrument and the conditions under which it is administered. The construction of one of the questionnaires has been described in some detail in order to illustrate how certain criteria which were demanded by these three assumptions were applied. If these criteria are judged to be adequate and the items of the questionnaire meet the criteria, then the instrument is one which is so constructed as to make possible the collection of valid evidence of the seven types of response to reading.

Valid evidence of these types of response, however, may not be given by the questionnaire even though its construction is judged to be satisfactory. Obviously, if such an instrument as Form 3.32 were administered as a "final examination" and the students informed that their grades or credits would be determined by their scores, we would not expect it to yield valid evidence of those students' responses to their

voluntary reading. The conditions which should attend the administration of one of these questionnaires are as follows: First, the teacher should understand the kinds of evidence the questionnaire is designed to give and should desire to secure this evidence. Second, the teacher should have a curriculum program which might be expected to bring about the development of the seven types of response. Third, the teacher should have developed a rapport with the students which will enable and encourage them to respond honestly to the questions. Fourth, the students should understand and accept the purpose of the administration of the questionnaire and the uses which are to be made of the results. This is merely to say that an evaluation instrument must be understood, must be relevant to the objectives and the curriculum, and must be accepted by the students as an opportunity to appraise themselves, if its use is to be of greatest value.

The assumption that students will respond honestly is a crucial one in these questionnaires, and unless it is fulfilled we cannot hope for valid evidence. In the construction of the questionnaire an attempt was made to select items which would not tempt students to be dishonest in their responses, and the directions were so phrased as to emphasize the desirability of answering as frankly and as honestly as possible. These were efforts to aid in securing honest responses. However, these efforts cannot be expected to make certain that the assumption will be fulfilled. The degree of rapport between teacher and students, students' previous experiences with "tests" and with the uses of test results, and students' concepts of the purposes of education and of the place of evaluation in education may determine to what extent the responses will be honest ones.

The questionnaire technique which is used in these instruments differs from the method of direct observation of students by a teacher only in that the student is both subject and observer rather than being merely the subject. One

method, then, of checking the honesty of a student's responses to the questionnaire would be to compare his responses with observations made by one or more adults of what he actually does and says. It should be possible for one familiar with the overt acts and verbal responses included in the questionnaire to compare his observations of some of these behaviors with student's responses. For example, a teacher might provide periods for "free-reading" and during those periods determine to what extent the student welcomes interruptions of his reading, reads various types of fiction and nonfiction, reads attentively, etc. Also, in conversation with a student, a teacher could secure evidence which would help her judge to what extent certain wishes and feelings expressed in his responses to the questionnaire were genuine. This is one method of validating responses to the questionnaire.

A somewhat different method which might be used would be to interview a student about his reading behaviors and in addition to asking him what he does, ask him for illustrations or examples of these behaviors. For example, a teacher who wished to know whether or not a student reads book reviews in current publications rather regularly probably could discover this without attempting to observe such reading directly. By asking him whether or not he ever read book reviews and, if his reply were *yes*, following this by asking in what publications he read them and what reviews he had read recently, and by giving him an opportunity to discuss some of these reviews, she could be reasonably certain of whether or not he actually did such reading. Such a procedure, of course, need not be an inquisition nor need it result in only an answer to the teacher's question. Reading guidance might be given as well as reading behaviors appraised in the same conversation.

Recognition of this method as a means of achieving reasonable certainty about what students actually do and say

leads to the possibility of constructing a paper and pencil instrument which would achieve a similar result. The student might be asked to respond on paper to questions about his reading behavior and then write out an illustration or an example of each behavior. The nature of the illustration or example presumably would be evidence which would tend to substantiate or refute his contention that he engaged in such behaviors. Let us for convenience call this a "direct form" of the questionnaire. The first page of such a direct form is reprinted below.

Name _____ Age _____ Sex _____
Grade _____ Instructor _____

This is not a "test" but an attempt to discover more about your reading interests. Obviously, no two persons have exactly the same reading interests; consequently *there are no "right" or "wrong" answers, as such, to these questions.*

Please answer each question as carefully and as honestly as you can. Mark your answer to each question by checking the space under *Yes*, *No*, or *Uncertain* at the right of the sheet. If your answer to a question is *Yes*, please give the additional information asked for in the question. If your answer is *No* or *Uncertain*, go on to the next question.

- | | Yes | No | Uncertain |
|--|-------|-------|-----------|
| 1. Do you have in mind one or two books which you would like to read?..... | _____ | _____ | _____ |
| If you do, please give the author and title of one: | | | |
| 2. Do you ever read adventure novels in your spare time?..... | _____ | _____ | _____ |
| If you do, please give the author and title of one which you have read: | | | |
| 3. Do you ever read essays, apart from school requirements? | _____ | _____ | _____ |

If you do, please give the author and title of one which you have read:

4. Is there any author whom you like so well that you would like to read any new book he might write?..... _____
If there is, please give his name and the title of one of his books which you have read:
5. Do you ever of your own accord read humorous stories or books of satire?.... _____
If you do, please give the author and title of one which you have read:
6. Do you ever read biography, apart from school requirements? _____
If you do, please give the author and title of one which you have read:

Such "direct forms" of the questionnaire have been used in studying the functioning of the Questionnaire on Voluntary Reading. The methods and the results of these studies will be reported in full in a forthcoming monograph. In brief, we find, for some classes, a relatively high relationship between responses on the Questionnaire on Voluntary Reading and on a direct form. These relationships, expressed as product-moment correlation coefficients, range from .38 to .79.⁸ Other types of studies which make use of interview techniques and of comparison of teachers' ratings of students with test scores will also be reported in the monograph. Similar studies of students' responses to the Novel and Drama Questionnaires have not been made; the presumption would

⁸ Fourteen such coefficients derived from a study of Form 3.32 are distributed as follows: .35 to .40, one; .45 to .50, one; .60 to .65, two; .65 to .70, three; .70 to .75, three; .75 to .80, four. The median of this distribution is .695.

be, since the basic technique is similar to that of Form 3.32, that such studies would yield results much like these. Tests 3.1, 3.7, and 3.8 were described as experimental instruments and the fact that they have not been studied has been mentioned.

Uses of the Instruments

Two major uses of the instruments described in this section may be pointed out: (1) To provide information about students which will aid in planning the school program and in guiding students; (2) To provide evidence on which can be based an appraisal of the progress of students and of the effectiveness of the school program. Before instruments such as the questionnaires described here are used, however, it is important for the teacher to examine the instruments carefully and to satisfy herself that they deal with behaviors which she regards as important. When such instruments are used, it is also important to recognize the limitations inherent in them and to supplement the evidence given by them with evidence gained from classroom observation and from other instruments. In interpreting scores on these instruments, it is important to consider the reliability data which are furnished in the Appendix and to use caution in making judgments based on differences in scores, either between individuals or groups.

The kinds of information given by these instruments have been described above. Such information as that given by the Questionnaire on Voluntary Reading should be of use to a teacher early in the school year to aid her in becoming acquainted with some of the reading behaviors of her students. For example, a teacher might profitably make use of the information that certain students or certain groups of students make very low "Appreciation" scores on the category "Likes to read." Assuming that a favorable attitude toward the reading of books is of some importance, either as an end in

itself or as a means to other ends, the teacher might plan special classroom experiences which would help these students to overcome the unfavorable attitude and to develop a favorable attitude toward books. In planning these experiences the question of why these students do not seem to like to read would necessarily be raised. In order to answer this question a number of hypotheses would have to be explored. Here the teacher would want to make use of evidence from other tests, such as tests of reading comprehension, from classroom observations made by other teachers, and from the school and home records of these students.

Such exploration of hypotheses might lead the teacher to give special attention to the reading behaviors of certain students as well as of the class as a whole. In planning reading experiences for individual students she also might find scores on the questionnaire helpful. For example, discovery of a student with a high "Appreciation" score on the category "Likes to read" but with relatively low "Appreciation" scores on the other categories might prompt the teacher to help the student discover and participate in such reactions as evaluating reading or relating it to life. Teachers have found that a conference early in the year with individual students which begins with the consideration of test scores may lead to an enthusiastic planning of individual programs of reading and other activities by the students themselves. In such conferences, of course, test scores should not be regarded as "marks" or judgments but instead as evidence which should be considered in planning the work of the year.

The second use is that of providing evidence on which appraisals may be based. Evidence of change from year to year in the status of individual students in their reactions to voluntary reading should be given by such an instrument as the Questionnaire on Voluntary Reading. This evidence should be useful to the student who wishes to make an appraisal of his achievement, to parents who wish to appraise

the progress of their children toward goals such as developing a favorable attitude toward voluntary reading, and to teachers who wish to appraise the success of their guidance and instruction in aiding students to cultivate some of these responses to reading. The appraisal of their own achievement by students is probably a necessary concomitant in any plan of promoting student as well as teacher planning of the educational program. Such appraisal, in turn, should stimulate further planning by both teacher and student. When the interest of parents in the success of their children demands more than a summarizing mark, a description of change in status as revealed by test scores should provide useful evidence to supplement anecdotal records or comments of the teacher. It is important, of course, for those who interpret these scores to others to make sure that changes in test scores are not mere chance fluctuations, but are "significant" differences, before interpreting them as such.

The role of other instruments in aiding the teacher in planning or in appraising her program should not be overlooked. Let us recall the three questions which members of the Committee on the Evaluation of Reading wished to be able to answer; namely, (1) How well does the student read? (2) What does the student read? and (3) How does the student react to his reading? An answer to the first question may be needed to help explain why a student does not read, of his own accord, or does not like to read. An answer to the second question may be needed to help explain why a student does not relate his reading to life. Thus in establishing hypotheses about the causes of certain students' difficulties in responding to reading it may be necessary to make use of several instruments which were designed to measure somewhat different behaviors. On the basis of such hypotheses, educational programs which are relevant to the particular needs of the student or group of students may be planned. In appraising the program it may be desirable to make use

of several instruments again in order to determine to what extent each of these behaviors has been modified. Consequently the use of such an instrument as the Questionnaire on Voluntary Reading may not be a sufficient evaluation procedure in itself. Those who wish to develop a more comprehensive plan of evaluation of reading behaviors should find the description of the instruments designed to help determine how a student reads and what he reads pertinent to their needs. These descriptions appear on pages 319 to 337.

THE EVALUATION OF THE APPRECIATION OF ART

The Committee on Evaluation in the Arts, composed of art teachers in the schools of the Eight-Year Study, listed as purposes of art teaching the following: (1) objectives pertaining to the development of sensitivity to art values, commonly called appreciation; (2) objectives related to the development of the ability to express certain types of experiences creatively; and (3) objectives related to emotional adjustment resulting from the release afforded by creative experience.

The evaluation of the first of these objectives—the development of sensitivity to art values—is the one with which the staff has been primarily concerned. Emotional adjustment can be fostered by means of well directed creative experience in the arts but the question of which are the particular types of emotional problems that can be solved, as well as the question of which kinds of creative experience offer a remedy for a particular emotional problem, is as yet not definitely answered.⁹ So it was felt that the primary consideration was the evaluation of sensitivity to art values and, although some attention was devoted to the emotional connotations, the results are not as yet sufficiently established to

⁹ The more important literature concerning this problem is cited in Levey, Harry, "A Theory Concerning Free Creation in the Inventive Arts," *Psychiatry*, III (May, 1940), p. 229 ff.

warrant extensive discussion. Furthermore, the area of personal and social adjustment was being explored separately (cf. Chapter VI); consequently only casual remarks on this aspect of the objective will be made in the following pages.

The problem of evaluating sensitivity to art values was further narrowed to include only the field of the visual arts. Here again it seemed unnecessary to duplicate work done in other areas. The evaluation of the appreciation of literature is discussed in the preceding section; other instruments of evaluation of appreciation in the field of the arts will be discussed on page 307. Thus the task became one of developing evaluation instruments which would appraise the students' sensitivity to art values in the field of the visual arts.

Ways of Getting Evidence and Exploration of Possible Criteria for a New Instrument

The first step in the study of the problem was to survey currently used methods of getting evidence regarding art experiences and art appreciation of students. Some of the methods which have been used to discover the development of the subject's knowledge regarding art—his *intellectual understanding* of art—include art questionnaires, art vocabulary tests, and similar instruments. These tests have attempted to appraise primarily the extent to which the student is familiar with art history and art techniques. Other tests have attempted to obtain an appraisal of the extent to which the student is able to apply certain rules of color-combination, balance, etc., in dealing with art objects. The success of the student on all of these tests seems to be chiefly dependent upon the extent to which he has mastered a body of factual knowledge which may be helpful in bringing about an esthetic experience.

Another approach to evaluation in the arts is through tests which attempt to measure the extent of the subject's *interest* in art and to discover in which sub-fields he has a special

interest. Still another method of gathering evidence regarding art experience has been to rely on a student's opinion about these experiences. His opinions may be stated in essay form or they may be expressed as responses to a checklist. More informal methods frequently employed by teachers include anecdotal records about student behavior, collections, descriptions, or photographs of creative work, and checklists filled out by teachers. The advantages and disadvantages of all these methods were reviewed in an attempt to set up criteria for an instrument designed to appraise responses to art values.

First of all, it was thought that tests of intellectual understanding, of mastery of specific areas of information, while useful where information is a part of the objective, would not necessarily contribute to an appraisal of the art sensitivity of the subject. It was recognized that a student may be sensitive to art values even though he has not mastered a body of specific information or rules. The converse seems also to be true; that is, a student may be familiar with the meaning of technical terms, the facts of art history, and so on, without being responsive to artistic values. It seemed desirable, therefore, that an instrument of appraisal should be so constructed that it would depend as little as possible upon the student's previously amassed information regarding art. The fact that it would be extremely difficult to eliminate this element entirely was also recognized.

Even though written statements about art experiences have the advantage of being highly personal and, therefore, may give insight into the nature of the individual's reaction, they too have one important disadvantage—they are frequently unfair to the student who is relatively lacking in the ability to state his reaction in words. It should be recognized that not all students who are capable of genuine and deep art experience have correspondingly well developed verbal abilities. It is very likely, for instance, that some students

who have very little verbal facility find a means of expression in art.¹⁰ Finally, there seem to be certain immediately visible qualities in an art object which are extremely difficult to translate into words, even for the verbally gifted person. Painting and prose are seldom mutually interchangeable as a means of expression. For these reasons it was thought desirable to have the instrument depend as little as possible upon verbal expression of subjective reactions. Since it was recognized that it would not be possible to eliminate the verbal element entirely, the aim was to reduce it to a minimum.

Records of behavior, anecdotal records, and collections of creative work, whereas they have the advantage of yielding evidence about the personal art experience of the individual, also have disadvantages. For instance, they do not provide a uniform basis for comparisons between students; also they apply only to the students who are productive in the studio; they fail if a student does not attend art classes.

In summary it might be said that there seemed to be a need for a new instrument which, as far as possible, would be constructed in such a way as to satisfy the following criteria: (1) that the results should not depend primarily upon a body of factual knowledge; (2) that the results should not depend upon the ability to express art experience verbally; (3) that the responses should permit a comparison of different students on a uniform basis; and (4) that the instrument should permit the evaluation of the responses both of students who are known to be artistically creative and of those who have not as yet exhibited such talents.

It was thought further that the instrument should attempt to get at the person's reaction to a work of art as a unit or as a whole, rather than at reactions to specific, separate elements of an object of art. It is doubtful whether one can get

¹⁰ Moreover, it seems as if adolescents especially are reluctant to state their problems openly and verbally. To them the less obvious way of expression by means of creation and participation in the arts is one of the main ways of dealing with these problems.

a valid indication of the capacity for esthetic experience evoked by an art object and what this object conveys, by asking a person to react separately to line, spatial arrangement, or color. Although this seems true for the evaluation of the esthetic experience as a whole, for the evaluation of certain aspects of esthetic capability a person's response to certain specifics of an art object is also needed. This is particularly true if the teacher wants to know at what particular stage of development the student's reactions to certain known features of art may be. Two additional criteria, then, seemed necessary. First, the instrument should allow the student to react to the art object in an esthetic way and permit a response to the work of art as a whole; that is, to have as complete an art experience as possible. Second, the instrument should contain a variety of elements and evoke specific responses so that the examination of these reactions of the student would permit an evaluation of his esthetic development with reference to these known elements.

Some Remarks on the Psychology of Art Appreciation

Before discussing in detail the specific assumptions underlying the development of the instrument, some further remarks concerning "art appreciation" should be made. Unfortunately the connotations of this term vary in different contexts and no definition is generally accepted. Sometimes the term is used in a rather narrow sense, covering only a passive act on the part of the beholder who in this context is compared with a piece of wax that bears the impression of a seal. A recent theory recognizes a great deal more activity on the part of the beholder who is supposed in the act of "empathy" to neglect his own personality and to live in the world of the work of art for the span of time during which he is in "empathy." A still more recent theory is that offered by "Gestalt" psychology. In dealing with these problems, from the point of view of this psychology, art appreciation is con-

sidered as a field phenomenon,¹¹ the field consisting of the beholder and the work of art. The act of art experience can take place—the field can be established—only if the spectator is willing to undergo the art experience. This willingness is a deliberate act on the part of the spectator, and art appreciation becomes an active rather than a passive reaction. In this connection it may be mentioned that for other and more elaborate reasons John Dewey¹² suggests that the term “art appreciation” may be discarded for the term “art experience,” and the latter term implies activity on the part of the beholder.

If art experience is conceived of as a field phenomenon, then the field will be strongly conditioned by the difference in the degree to which any one of the main elements constituting the field governs it. One extreme would be a situation in which the work of art dominates the field, a situation close to the one mentioned above in the example of the seal on wax. Fortunately this situation never occurs because even the most passive spectator is still a personality with a particular background, particular education, particular opinions and feelings about art, which, even though he may be unaware of them, will influence the field. The other extreme would be a situation in which the spectator dominates the field and is not touched at all by the work of art. It might be said that he is in a situation in which he is confronted with a work of art which he sees but does not experience. The ideal situation is a playing back and forth within the realm of the field, the spectator becoming more and more incited to bring new facets of his personality into play, and in turn becoming more aware of new facets of the work of art. Spectator and work of art may be said to be communicating with one another, a communication which is strongly conditioned

¹¹ See Koffka, “Psychology of Art,” *Bryn Mawr Symposium on Art*, p. 224 ff.

¹² See, for example, John Dewey's recent volume, *Art as Experience*.

by the nature of both of them. The importance of the personality of the spectator, his experience, and his emotional predispositions, may be corroborated by the well-known fact that at different times in life we experience works of art in different ways.

It thus becomes apparent that when one learns which aspects of a work of art are important for the art experience of an individual, access has been gained not only to his particular way of experiencing art, but also to his personality. It is even more important to ascertain which works of art induce a spectator to have this personal experience—to learn which works of art incite him to establish this field phenomenon called art experience. Moreover it is of interest to find out which works of art “leave him cold,” because they are to him void of meaning, or because they seem too unimportant to him to induce the amount of interest necessary for experiencing them. Again this will shed some light not only on the character of the spectator’s art experience but also on his personality. If something about the personality of a student can be learned by studying the environment which he creates for himself, by exploring the kinds of persons he prefers to be with, or the kinds of persons that he avoids, then the type of pictures with which a person does or does not “communicate” may be indicative not only of his art experience, but also of his personality. Finally, one wants to learn whether or not a person actually *prefers* the works of art with which he is able to communicate.

The possible bearings of art experience on creativity in the field of art deserve comment. Obviously only the person who is able to experience in an esthetic way objects and events of the outer world, art objects as well as others, is able to express these esthetic experiences creatively. It was assumed that artists perhaps more than others are capable of having esthetic experiences with objects not yet molded into esthetic wholes. Moreover, during the process of expression or creation

the emerging product has to be evaluated by the artist in terms of his esthetic perception, in terms of the evolving product's suitability to induce or evoke art experiences in an ideal beholder.¹³ Therefore it is to be expected that the art experience of the artist would not be essentially different, but only more highly and more intricately developed when compared with the art experience of the non-artist. It might also be expected that persons whose art experience is highly developed need not be or become artists, either because of lack of skills or because of other reasons. On the other hand, one would expect the artist's art experience to be of the highest quality. Moreover, the person who demonstrates a high degree of esthetic sensitivity in relation to the extent of his art experience may be a latent or future artist.

Although the above remarks are not adequate for covering the topic with which they deal, it seemed desirable to clarify to a certain extent the theoretical framework underlying the assumptions on which the development of the new instrument was based. These assumptions will now be discussed.

DEVELOPMENT OF THE INSTRUMENT

Basic Assumptions

The basic assumption of the new instrument to be described in the following pages is that it is possible to understand the nature of and degree to which the art experience of an individual is developed by ascertaining the degree to which he is able to see and appreciate significant similarities and differences in art objects. "The reaction of the artist is colored by all sorts of . . . associations and feeling, of which he is naturally unaware, but which affect profoundly the form taken by the work of art and which have the power to stir up corresponding . . . feelings in the spectator. It is the

¹³ It is not implied that the artist tries to "please" the general public, but that his efforts are concentrated on organizing his creation in such a way that it may be suitable for conveying his esthetic message.

fact that the works of art act as a transmitting medium between the artist's . . . nature and our own that gives it its peculiar, and as we may say 'magic' power over us. It is 'magic' because the effect on our feelings often transcends what we can explain by our conscious experience."¹⁴ If the reactions of the artist, colored and conditioned by his personal associations and feelings and embodied in his work of art, have stirred the spectator to corresponding—even though not necessarily identical—feelings, the artist (or actually the work of art) and the spectator may be said to be communicating with one another. This communication is possible if the spectator has been able to establish an esthetic field including himself and the art object. When this happens, we may say that he really is able to "appreciate" the work of art, that he is "sensitive" to its artistic qualities.

The deeper the art experience of the subject is, the more he responds to the personality of the artist as revealed in the work of art, the specific way in which the artist rendered his subject matter, the cultural background of the work of art, the importance of the media chosen, the particular way they are used, etc. The quality of his art experience is developed to an even higher degree if he is responsive in this way to different works by the same artist, though the subject matters and other more superficial qualities (such as the size of a picture) may differ from one work to the next.

A first assumption, then, may be that art sensitivity is revealed by the *degree* to which a student responds to the visible similarities existing in certain works of art created by the same artist. As a matter of fact, the degree to which these similarities can be seen and the degree to which a subject can reasonably be expected to respond to the affinity existing between the objects created by one artist will depend on many factors. Some of these factors, such as the particular

¹⁴ Fry, Roger, *Art History as an Academic Study*, p. 13 in his "Last Lectures."

selection of works of art being viewed and the context or conditions under which these works are seen, are of outstanding importance. Unless these are properly controlled, the assumption may become invalid.

A second assumption is that the *nature* of a student's art experience may be revealed by the *kinds* of similarities to which he is or is not responsive. He may be responsive to the similarities existing between works of art seen as wholes, to the affinity mentioned before, or he may be responsive only, or chiefly, to similarities in color, mood, or spatial arrangement. If enough opportunities are given to a student to select similarities, his pattern of reaction may be open to examination. This may also be said to be true in a negative sense; that is, it may be characteristic of a student *not* to see, or to be unresponsive to certain kinds of similarities.

A third assumption is that a student whose appreciation is well developed will have a certain definite emotional reaction to art objects. He will like works of art which make use of the qualities he is responsive to; he will dislike art objects which make use of qualities that do not appeal to him. He will neither like nor dislike art objects which "leave him cold," which "do not convey any meaning," i.e., art objects which seem uninteresting either way.

Construction of the Instrument

The construction of the instrument, "Finding Pairs of Pictures," was based largely upon the three assumptions discussed above. The instrument had to provide evidence as to the degree to which, and the way in which, students respond to the affinities existing between works of art; it had to provide evidence concerning the kinds of similarities to which they are responsive or unresponsive; and it had to reveal the art objects, or qualities of art objects, to which they have a definite emotional reaction.

According to the first and second assumption, it is possible

to understand the nature and degree of art experience of an individual by ascertaining the degree to which he is able to see and appreciate important similarities and differences in art objects. It was thought that this might be tested most appropriately by presenting students with examples of art objects and asking them to pair them, and then examining the results to see what inferences might be drawn. The third assumption—that students will have an emotional reaction to art objects which make use of qualities to which they are responsive—could be tested by asking the students to select certain examples which they liked or disliked for certain reasons, and examining these choices to see whether or not they corroborated hypotheses raised by the examination of the pairings.

In constructing the instrument it was impossible to present a great variety of art objects at one time and hence for practical reasons a restriction to one field of the visual arts was necessary. A decision was made to begin with the construction of a test covering the field of painting. This field was selected for two reasons: (1) it is more complex than some of the minor arts, and (2) students are usually more familiar with it than with sculpture, architecture, or with the minor arts. There is also the possibility that the response to certain subtle values in paintings may be a valid indication of esthetic response to the same values when they appear in other fields of the visual arts. For instance, one would expect a person whose response to color combinations in paintings is well developed to be able to apply the same discrimination in dealing with textiles, etc. This will have to be tested in future studies, however.¹⁵

The next problem after limiting the field to that of paint-

¹⁵ It is realized that ideally an evaluation of art experiences should cover all the fields of the visual arts, and it is thought that tests based on similar principles but covering other areas, such as sculpture, architecture, and the minor arts, can and should be developed.

ing was that of setting up criteria for the selection of the paintings to be used for the pairing.

In the first place, the pictures had to be selected in such a way that there would be an optimum chance for creating an esthetic mood. It was desirable that everything endangering this mood should be avoided as far as possible. It was necessary to exclude pictures evoking too strong effects and pictures evoking extra-esthetic deliberations, if not very special reasons recommended using them. Thus, it was decided that certain subject-matter fields could not be used because they dominated the students' interest too strongly. For instance, a picture such as "Washington Crossing the Delaware" could not be used because primarily it evokes patriotic feelings or historical deliberations, rather than "purely esthetic" feelings. Because it was found in preliminary studies that some students have difficulty in pairing pictures from widely different subject-matter fields, it was thought desirable to limit the subject-matter somewhat in order to provide a maximum opportunity for pairing.

It was also felt that students brought up in the tradition of appreciation for the old masters and students whose main interest is concentrated on modern art should, in taking the test, have about the same opportunities to reveal sensitivity to art values. Therefore, it was necessary to exercise care in order that the selection not be dominated by one group or the other.

Most important of all, however, was the selection of examples which could be legitimately paired; that is, examples containing affinities which can be recognized by students. It was realized that the similarities between the paintings of a single artist may not always be greater than the similarity of certain elements of one of his paintings to the same elements in a painting by another artist. Care had to be exercised to remove as many of these potential sources of confusion as possible. To assure this point, it was decided that the selec-

tion of paintings to be paired should be made on a strictly empirical basis. In line with this, a series of experiments was made with a group of 60 high school students who were chiefly in the ninth, tenth, and eleventh grades. Several hundred reproductions of paintings were presented to them in groups of about 40 paintings, and a careful record of their responses was kept. Pictures which were not used at all by these students for pairing were discarded at once. Pictures which were paired with pictures by other artists in more than 25 per cent of the total number of times they were used were also excluded from further experiments. The remainder of the pictures which had been paired with those by another artist were dealt with in a manner which can be best described by giving an example of what actually happened.

One group of paintings presented to the students of the experimental group contained among other pictures two paintings by Picasso, "The Absinth-drinker" and "The Guitarist"; several paintings by El Greco, among them the "View of Toledo"; and several paintings by Corot, among them "Paysage."

In more than 25 per cent of the times any one of the two paintings by Picasso was used for the purpose of pairing, it was paired with the other painting by Picasso. Therefore, the experiments with these two Picassos were continued. Suppose one student paired "The Absinth-drinker" by Picasso with the "View of Toledo," while another student paired the same picture with the "Paysage" by Corot. This suggested that in a complicated situation, when many elements from which to choose are offered, it is difficult for some students to respond to the affinity existing between these two Picassos. Therefore, a less complicated experimental situation was set up. To students four pictures were presented, the two Picassos, the Corot, and the El Greco, and they were asked to find the picture closest to "The Absinth-drinker." In other words, this time they did not have to select one out of 39

pictures, but one out of three. Unless at least 90 per cent of the students of the group selected the other painting by Picasso as the best choice, the painting "The Absinth-drinker" would have been excluded from future experiments.

This procedure was followed with all paintings for which some doubt existed about whether or not they ought to be excluded from the test. The purpose of this procedure was to make sure that the pre-supposed affinity existing between paintings by the same artist actually exists for students of this age level and cultural background. By selecting the sample in this way it was hoped that as far as possible no standards would be imposed on the students which might be outside of their particular experience or alien to the orbit of their esthetic perception.

In selecting the material for the instrument, then, the samples were restricted to the field of painting; pictures were selected in such a way as to provide an optimum chance for creating an esthetic mood; pictures which might prove too distracting were avoided; examples were restricted to a few subject-matter fields; care was taken to provide examples of the works of old and modern masters; and as far as possible only those pictures by any one artist were chosen which, according to preliminary experiments, had similarities which students are able to recognize as such.

DESCRIPTION OF THE TEST

As finally developed, the instrument consists of a picture sheet, a set of instructions to the student, and an answer sheet.

The Picture Sheet

The *picture sheet* consists of a piece of cardboard, approximately 24" x 44" in size, on which 40 colored postcards are mounted. These are copies of more or less well-known paintings ranging in periods represented from the Italian and

German Renaissance to modern and contemporary art. Dutch, Spanish XVIIth century, and French XIXth century paintings are included. Portraits, landscapes, and still-lives are represented. The copies used are of the best available quality, chiefly Jaffe prints, and they have been arranged on the cardboard in such a way that the whole set makes in general a pleasant appeal. Particular effort has been made to avoid having one painting interfere with the appreciation of another next to it. No titles or names of artists are given, but each painting is marked with a number for identification.¹⁶

The Instructions

The *instructions* presented to the students are so stated as to reassure them that the test is not based on any particular notions about art or painting, periods or painters. They are told that it is not expected that the art appreciation of an individual ought to conform to any fixed standards. Efforts are made to convince them that art appreciation is something very personal, different from one person to the next. Therefore it is carefully pointed out that there are no "right" or "wrong" ways of going about taking the test.

Deliberate efforts are made to avoid as far as possible restrictions which might limit the response, or create an atmosphere of examination. Thus students are told that no time limit is set, and, even though according to experience the student's ability to find pairs is usually exhausted after about 45 minutes, it is recommended that teachers allow students to use as much time as they wish in taking the test.

Other limitations of the response would be to ask the students to use every picture, or to find a prescribed number of pairs. In order to avoid this type of restriction it is pointed out to the students that they are not required to use every one of the pictures, that they may use one picture several

¹⁶ For the list of paintings used see the Appendix.

times for the purpose of pairing, another one not at all. A certain freedom is given to the students in determining the number of responses they wish to make. During the preliminary experiments the students were not told to select any particular number of pairs; nevertheless the great majority selected between 20 and 30 pairs. Experience to date has shown that nearly all students are able to find about 20 pairs and that after about 23 pairs most of the students stop working. On the basis of this experience, the students are asked to find, if possible, at least 20 pairs, but not more than 30 pairs.

The instructions suggest the selection of pairs of pictures which have important artistic features in common. As examples of such features style of painting, use of colors, design, mood, the way in which objects are painted, are mentioned. Since experience demonstrated that most students show a tendency to rely too strongly in their pairing on the similarity of subject matter, they are warned that: "If a subject matter in two pictures is the same (such as flowers), but if each of them is painted in a different way, then this similarity of subject matter does not seem to be an important reason for pairing them. It might be better to put one of these paintings of flowers together with a portrait, or a landscape in which the colors and the design, the style and the mood are very much like those used in the painting of flowers."

The Answer Sheet

The students are asked to indicate their choices of pairs and their preferences and dislikes of pictures on an *answer sheet* prepared for this purpose. The answer sheet consists of two parts and contains in its first part, in addition to the usual identifying data, spaces in which the students can indicate their selections of pairs by writing the numbers of the two paintings which according to their opinion have important artistic features in common. This part is arranged as follows:

1. No. _____ and No. _____ make a pair. 2. No. _____ and No. _____ make a pair, and so on up to 30.

The second part of the answer sheet is arranged as follows:

Now that you have studied all of the pictures, give some general information as to your personal preferences and dislikes.

1. Please give the numbers of 1, 2, or 3 pictures which you like *best*:

The numbers of these pictures are _____

I like these pictures best because _____

2. The picture I like best for the mood is picture number _____

3. The picture I like best for the colors is picture number _____

4. Please give the numbers of 1, 2, or 3 pictures which you like *least*:

The numbers of these pictures are _____

I like these pictures least because _____

5. The picture I like least for the mood is picture number _____

6. The picture I like least for the colors is picture number _____

THE TEST INTERPRETATION

The Scoring

The basis for the scoring is the number of pairs of pictures painted by the same artist which a student is able to find. Pairs of pictures painted by the same artist will, for convenience, be called "S" pairs. The pictures used permit the selection of as many as 43 different "S" pairs.

One of the "S" pairs consists, for instance, of the pictures No. 1 and No. 24 (see list of paintings in Appendix). Both are paintings by Picasso, painted in his so-called "blue" period. The color scheme used in both paintings is very similar, and no other painting is included in the test which has

an analogous color scheme. Both paintings are representative of a certain period of painting. The particular flow of lines, the sad mood expressed in them, the way in which the subject is rendered, and many other features can be found only in these two paintings in the present set. If a student selects this pair we may assume that he is responsive to several of the artistic similarities these two paintings have in common, and, moreover, that he probably is responsive to the affinity existing between the two pictures as a whole.

A copy of a score sheet is reproduced in the Appendix. Three scores are given in per cents—the “S” pairs a student was able to find; the ratio of the number of “S” pairs to the number of attempts; and the number of artists, expressed as a per cent of the total number of artists, whose paintings the student was able to pair in an “S” way.

Of these three elements the most important and most informative is the second. The first score obviously is conditioned by the willingness of a student to select many pairs; by pure chance a student who selects 30 pairs ought to find more “S” pairs than one who selects only 20 pairs. Therefore, the per cent of “S” pairs has to be interpreted in the light of the number of attempts the student made; this is facilitated by the second score. The score on number of “S” pairs is recorded because if two students, for instance, have about the same score in “Ratio,” the one with the higher score in “S” pairs obviously has given a better performance.

The score on “Number of artists” is mainly of descriptive character and may be used for the purpose of ranking students only if the score on “Ratio” as well as on “S” pairs is nearly the same for two students. Actually this score is separated into subscores and the record on the right side of the score sheet indicates those artists whose paintings a student was able to pair in an “S” way.

If a student paired only or primarily old masters in an “S” way, one may infer that this is the realm of his main interests.

If a student found a great many "S" pairs by using only paintings by one or two masters it might be that he is for one reason or another very well acquainted with just these paintings, and it may be inferred that the range of his understanding is smaller than is indicated by the score on "S" pairs and "Ratio." The statements regarding preferences and dislikes are not recorded on the score sheet because so far no way of treating them numerically has been found.

The scores indicate to what degree the student's appreciation of the 40 paintings included in the test is developed as compared with other members of his group. They indicate roughly whether his appreciation of modern or old masters, of portraits or still-lives, is developed to about the same degree, or is unevenly developed in any one of these areas. By means of the scores alone it is not possible to ascertain whether a student has native artistic ability or only an intellectual understanding of the field. A high score may be due to native ability or it may be due to the special background of the student. Familiarity with art, frequent visits to museums, and the like, influence the score in the same way as creative work in the arts or native abilities influence them. Nevertheless, the rough score seems to indicate fairly accurately where a student stands within his group with respect to the degree to which his art experience is developed. If one wishes to know more about a student, his individual responses must be examined, since the answer sheet furnishes information which is not reported on the score sheet. The method of obtaining this is to make an interpretation of the data recorded on the answer sheet.

The main assumption underlying this interpretation is: everything that the subject does is important and he does not do anything without valid reasons. The basis for a given reaction of a student may or may not be a genuine esthetic response to an art experience; nevertheless, in interpreting the results of the test, one ought to be able to answer certain

questions. For example: What were the main artistic features to which the student responded? What are the artistic features to which he seems to be unresponsive? What might have been the reasons preventing him from making an esthetic response to the art objects presented to him? Or, approaching it in another way, one might ask what might be the reasons within a student's personality which made him respond to certain works of art, or particular art elements, and not to others. To answer these questions, the study of pairs consisting of two pictures painted by different artists is as important as the study of the so-called "S" pairs. The former pairs may be called "D" pairs.

A "D" pair which is occasionally selected by some students consists of No. 1 and No. 35. Both paintings make use of greenish colors, but their use, the way they are blended, and their meaning within the context of the painting is quite different in these two paintings. The mood expressed in both paintings is of a more or less introspective quality, enforced by the cold colors in which both are painted. The quality of this introspectivity is different, however. The mood of No. 1 may be described as being sad and withdrawn, whereas the mood of No. 35 is one of religious exaltedness. The style in which these pictures are painted is different, but there may still be recognized in both a common "Spanish" element. The selection of this "D" pair may be accepted as indicating that the subject who selected it was responsive to the general color used in these paintings, even though he was not responsive to the different ways in which these greenish colors are blended. He probably was responsive to the general mood of introspectivity permeating both pictures, without being responsive to the important difference in mood which can be recognized. The student may have been responsive to the "Spanish" element common to Nos. 1 and 35 without being responsive to the difference in the style.

Many more inferences pertinent to the student's art experi-

ence and in this way pertinent to his response to art as well as to his personality might be drawn from the fact that he selected this particular pair. Great caution has to be exercised *not to consider valid an inference based on the study of any one pair*. The selection of any particular pair can have a quite different meaning when occurring in different contexts. Any one response to this instrument has to be interpreted in the light of the possible meaning of all other evidence which can be obtained through a study of all of the responses of the subject to the test. In this connection, as has been mentioned before, not only what a student does is of importance, but also what he avoided, or missed doing, has significance. Pairs he selected not only have to be studied in the context of all other pairs, but they have to be studied in their sequence, and in the light of the pairs the student failed to select. First we have to consider which are the pictures he likes and dislikes; these data in turn will shed light on the pairs selected because students tend, in their pairings, to make different uses of preferred and of disliked pictures.

When the present study of this instrument is concluded, all pairs which have been used to a considerable extent and which seem to be significant either for the art experience or the personality of a student, will be listed, each with the inferences which suggest themselves in connection with the use or non-use of the pair. Once this list is available the interpreter will have to integrate into a consistent picture the meaning of the pairs which a student has selected plus the meaning of the non-use by this student of pairs *commonly used*. This integration will have to be achieved through considerations of the meaning of the preference or the dislike of any one of the 40 pictures.

This task will be less difficult than it appears, because we can restrict the investigation of the student's responses to the areas in which he differs from the group. The "Ratio" score which a student obtains places him in a certain section of his

group. The importance for the interpretation of his selection of any one pair depends upon the extent to which it is similar in difficulty to the other pairs which he has used. An example might clarify this somewhat. If, for instance, a student who is in the lowest quarter of his class in his "Ratio" score selects an "S" pair which has been found by only one or two other students who are among those receiving the highest scores on "Ratio," this pair becomes very significant for the interpretation. It becomes significant because one would expect a student with a low "Ratio" score to be able to find only the most obvious pairs, that is, only the pairs which have also been selected by a large portion of the group. The opposite is also true—if a student is in the highest quarter of the class in his score on "Ratio," and one finds that there are pairs selected by a large portion of the group which he has missed or avoided using, these pairs become significant for the interpretation.

It is evident that a student's responses must always be examined against the background of the group and the way in which the members of the group have reacted to the test problems. This is not only true of the particular group in which the student is working but it is also true of large age, sex, and cultural groups. The study of these larger group differences will provide important material for future investigations.

As a basis for the test interpretation, the following information is therefore needed:

1. *An analysis of how often any pair has been used by the other members of the group.* This analysis will make possible a decision as to the degree of significance which might be attached to the selection of a pair. The kind of inferences which can be drawn if a pair has been selected has been indicated on pages 293-296. Here we may add some of the inferences

which might be drawn if a student does not select a particular pair. The "D" pair mentioned above consisting of Nos. 1 and 35 may again be used as an example. Assuming that this "D" pair has been commonly used by the other members of the group, some of the reasons for the avoidance of this pair might then be: a better developed discrimination for the importance of color shades, for differences in style, and a lesser degree of responsiveness to introspectivity.

2. *Knowledge of the average number of times any single one of the pictures has been used.* Continuing the example, we should like to know whether the student used pictures expressing an introspective mood less often than the average. If that is true, then the avoidance of Pair 1-35 might not be due to a higher discrimination, but may be due to a lack of interest in paintings expressing an introspective mood. In this connection it may be added that the use of a picture more often than the group average usually indicates that the student's interest centers around this picture. This interest need not always be of a positive nature. A repetition of a pair may also indicate a concentration of interest.
3. *A comparison of the preferences or dislikes of a student with the preferences and dislikes of his fellow students.* In continuation of the example mentioned above, we may say that if this student states that he prefers introspective pictures, or pictures making use of dark, greenish, or cold colors, we can be reasonably sure that he avoided the selection of Pair 1-35 for esthetic reasons. On the other hand, if he dislikes this type of painting, or is not at all interested in it, the avoidance of Pair 1-35 becomes less important as far as the evaluation of his discrimination for

artistic values is concerned. As another indication of his avoidance of introspective tendencies it will still be important for evaluating his personality.

4. *A study of the sequence of pairs.* Study of the meaning of the sequence of pairings has been very fruitful. For purposes of illustration of the kinds of insights this permits, the following illustration may be given. Certain very obvious pairs tend to appear in the very beginning of the test. A student who begins with seldom used pairs seems to be one whose art experience is different from that of others in the group. To begin by indicating pairs consisting of portraits is usual. To begin with a pair consisting of still-lives suggests either a person very much interested in this subject matter or a student who is reserved at first in establishing positive relations with his fellow men, or both.

It can be seen that the interpretations would be greatly facilitated if they could be made on the basis of a fairly large collection of data on the way in which members of different groups respond—the ways in which they pair the pictures; the pictures they like and dislike; and the sequences in which pictures are used. Thus far it has not been found possible to achieve this.

HOW TO ADMINISTER THE TEST

In accordance with our general conception of art experience, it is important that a spirit of freedom prevail during the time the test is taken in order that an esthetic mood may be created and maintained. It is best to have every student work with a separate picture sheet. However, two or three students may work together on one picture sheet. Although care should be taken that they do not unduly influence one another, nevertheless explicit prohibitions not to discuss the

test should be avoided. Some free discussion makes the selection of pairs much more interesting. Anything that can contribute to the students' feeling at ease should be done. Thus, they should be allowed to stand up and move around so that they may see the pictures better, etc.

RELIABILITY AND VALIDITY

Reliability

On *a priori* grounds it seems reasonable to believe that it is more difficult to secure an adequate sample of pictorial art (a field in which reactions may be strongly influenced by the emotions) than it is to achieve an adequate sampling of information within a restricted subject-matter area. Because sampling affects reliability, a reliability coefficient which would not be considered very high for an information test may be the highest reliability coefficient which can be expected on an art test of the type described.

Meier and Seashore, for example, state that "with tests based upon concrete learning accomplishment a higher reliability is expected than one testing complex mental functions. With the latter kind, a coefficient of reliability of .80 is regarded as about as high as can reasonably be expected, because of the uncertainty of knowing exactly what factors operate in the person's total reaction. With a test of capacity the opportunity for chance factors to control the final result is increased, hence a somewhat greater allowance must be made for them."¹⁷

Two reliability studies of the instrument under discussion here were made, the first based on the split-half method, the second on a comparable test form. The reliability coefficients estimated by correlating the halves of the test and applying the Spearman-Brown prophecy formula, based on the test results of 145 twelfth-grade high school girls and boys, are as follows: for the scores on "S" pairs, the coefficient is 0.57;

¹⁷ Art Judgment Test, Examiners Manual, p. 21.

for "Number of artists," it is 0.58. Since the ratio score for each half cannot be added to get the "Ratio" score for the entire test, we cannot give a statistical estimation of the reliability of the ratio score based on the split-half method. Therefore, a somewhat comparable form consisting of 49 other paintings in place of the original 40 paintings was developed. These 49 paintings are not as well known as the ones used in the original form, but they cover the same periods.¹⁸ The students who took both tests were not homogeneous in their art experience. The group consisted of 27 senior high school students and 38 college students. It may be expected that the results will be somewhat better if the experiment is repeated with a larger group. The second form was taken shortly after the first test was taken, either after a lapse of several hours, or within one or two days following. The reliability coefficients based on the intercorrelations of the two forms are 0.58 for "S" pairs, 0.77 for "Ratio," 0.54 for "Number of Artists."

As has been mentioned before, the most important score is the one on "Ratio." According to the directions of the test, which give great liberty to the students in selecting many or few pairs, and in using the paintings of many or few artists, it was not to be expected that the reliability coefficient of the scores on "S" pairs and on "Number of Artists" would be very high.

Validity

Validity studies are still in progress. Such evidence of validity as has been collected will be presented here, with the reservations which must accompany data which are incomplete. It was thought that the validity of this test might

¹⁸ For a list of these paintings use the Appendix. Some of the pictures in the comparable form furnished such interesting and important information that they ought to be included in a future form of the test in place of some of the pictures originally used which were less successful in yielding information.

be established if the following assumptions can be substantiated by the evidence:

1. The test measures some ability which is not a function of the particular pictures used in the test. The correlation between two tests which use quite different pictures would seem to indicate that the test is measuring an ability which is not dependent upon the particular pictures which make up the test, but an ability which does operate within a wide variety of pictures.
2. Subjects are responsive to one or more of the basic qualities of art, but are responsive in different degrees. This would seem to be supported by the fact that the lowest score on number of "S" pairs which any student made is higher than a chance score would be, and there is a considerable range—from 17 per cent to 100 per cent—in the scores of the subjects.¹⁹
3. The development of visual sensitivity or of art abilities need not correspond to the development of intellectual abilities as measured by the usual intelligence tests. In the case of one school, the results of this test were compared with the results of intelligence tests giving a correlation coefficient of approximately zero.
4. Since art is something which can be taught, at least up to a certain degree, the general level of a group of art students ought to be higher than the level of a comparable group without art training. The median score of a group of art students has been found to be higher than the median score of any unselected group. The groups are small, however, and no controlled experiments have been set up to indicate whether or not a further selective factor of ability or interest has

¹⁹ See table on p. 304. By mere chance a student might be expected to get a score of 5 per cent or less.

been operating to produce the results which we have at present. It would be desirable to repeat this study with larger groups of students and to compare the results with the results of control groups. (See table on page 304.)

5. Students with native ability should give a better performance on the test than students without native ability. Within a group without art training, therefore, there should be some students who, due to their native ability, perform as well as students with art training. This would seem to be corroborated by the fact that in the eighth grade the highest score made by any student is as high as the lowest score made by any student in the master class in painting in an art academy. This latter group is composed of students who intend to become professional artists. As can be seen in the table on the next page there is considerable overlapping in the ranges of scores of different groups. The weight which each factor, ability and training, contributes to the scores will have to be determined by a controlled experiment.
6. The student reveals the nature of his appreciation of art and some elements of his personality structure by his choices of pairs and by his preferences for pictures. Evidence for this assumption is encouraging though not conclusive. Unfortunately, many of the evaluations of the interpretations have been made in verbal rather than numerical form. It is impossible at this point to print them in full, or to ascribe numerical values to these evaluations.²⁰ In four schools, however, teachers were asked to select a number of students with whom they were very familiar. The test results of these students were in-

²⁰ A monograph in preparation will include more extensive discussion of similar studies.

No. of Students	Type of School	Scores in Per Cent
17	8th Grade, New England Boarding School	
10	12th Grade, New England Boarding School	
16	12th Grade, New York Private School	
69	12th Grade, Mid-Western Public School	
81	12th Grade, Western Public School	
26	Freshmen, Art Academy	
15	Freshmen, Mid-Western Teachers College	
14	Advanced Art Students in Mid-Western Teachers College	
11	Graduate Students in Fine Arts in an Eastern University	
7	Students of a Master Class in Painting in an Art Academy	

No. of Students	Type of School	Scores in Per Cent
17	8th Grade, New England Boarding School	
10	12th Grade, New England Boarding School	
16	12th Grade, New York Private School	
69	12th Grade, Mid-Western Public School	
81	12th Grade, Western Public School	
26	Freshmen, Art Academy	
15	Freshmen, Mid-Western Teachers College	
14	Advanced Art Students in Mid-Western Teachers College	
11	Graduate Students in Fine Arts in an Eastern University	
7	Students of a Master Class in Painting in an Art Academy	

terpreted and the teachers rated these interpretations on a five-point scale ranging from "very good" to "poor." The intermediate ratings were classified as "generally accurate," "possibly accurate, but insignificant," and "of doubtful value."

In one school the teachers selected 17 students. Of the interpretations of the test results of these students, nine received the highest rating—"very good." Four descriptions received the next rating—"generally accurate," and four the middle rating—"possibly accurate, but insignificant." No cases were placed in the "of doubtful value" or "poor" columns. The teachers were also asked to indicate any "gross inconsistencies or errors." They found none, and stated further that in no instance was there failure to designate at least one important characteristic of the student.²¹

In another school the descriptions were not only rated on the five-point scale. For some descriptions the teachers used ratings composed of two of the five points of the scale, indicating in this way that one part of the description seemed to deserve one rating, another part of it another rating. Thirty-three students were described; of these descriptions 16 were rated as "very good," four as partly "very good" and partly "generally accurate." Two were rated as "generally accurate," one as partly "very good," partly "possibly accurate, but insignificant." Three were rated as partly "very good" and partly "poor," two as "of doubtful value," none as "poor." Five descriptions were rated with different combinations of the five values of the rating scale.²²

The test results of 27 students of an art academy were interpreted and the faculty of the department of painting was asked to rate these interpretations on the same five-point

²¹ This study was conducted at George School, Bucks County, Pennsylvania.

²² This study was conducted at the Cambridge School, Cambridge, Massachusetts.

scale. Eighteen cases received the highest rating, six the next, none the middle rating, one was rated of doubtful value, and two received the lowest rating.²³

Finally, 31 students of a teacher-training institution were tested and the results interpreted. The descriptions of these students were rated by the art faculty, and 22 descriptions were rated as "very good," six as "generally accurate," two as "possibly accurate but insignificant," one as "of doubtful value," none received the lowest rating. It was added that "almost without exception the essential qualities of the students" were "clearly" mentioned in the descriptions.²⁴

The validity studies conducted at these four institutions are summarized in the table on page 534. According to this table, approximately 60 per cent of the descriptions were rated "very good," and approximately 81 per cent of the descriptions were considered as being satisfactory (either very good, or generally accurate, or of an intermediate value between these two). Approximately 10 per cent of the descriptions were considered as being unsatisfactory (either of doubtful value, or poor, or intermediate values between these two). Only 2 per cent of the descriptions were rated as being definitely of poor quality. It is hoped that this discussion will indicate the direction of the work on validity, both past and future, and the extent to which the evidence, however meager, supports the original assumptions.

FUTURE USE OF THE TEST

The study of this test has not matured to a point where it is possible to present scientifically dependable conclusions about how such an instrument can be used most efficiently. However, it does seem that the instrument may be used for the purpose of counseling in so far as it may be possible to

²³ This study was conducted at Cranbrook Academy of Art, Bloomfield Hills, Michigan.

²⁴ This study was conducted at State Teachers College, Milwaukee, Wisconsin.

decide where a particular student stands when compared with his peers as far as his response to paintings are concerned. Moreover, it may be possible to use this instrument to ascertain changes in the performances of students, or groups of students, after they have taken art courses. For these two purposes the scores seem to furnish valuable evidence.

This instrument can be used much more efficiently if individual interpretations are made. When this is done, it seems possible, by means of the instrument, to get evidence about the specific art abilities of a student as well as about some of the features of his personality. It will be possible to discover some of the areas where he needs special help. By repetition of the test, it will be possible to discover the areas in which he has changed and those in which he remained on the same level as before.

Finally, even at the present stage of development, the test furnishes some insights regarding the way in which art experience is tied up with personality structure. More extended studies will enlarge our understanding of important art-psychological questions, such as the ways in which art experience varies with different age and sex groups, different cultural groups, and groups from different socio-economic levels. Information as to the particular way in which the individual experiences are combined with information about the differences in the reactions of different groups should have implications for the teaching of art.

OTHER INSTRUMENTS

Several other instruments to reveal the ways in which students respond to art experiences were developed experimentally but were not studied as carefully as the one just described. One of these was called Seven Modern Paintings (Form 3.9). A committee of art teachers selected seven ex-

cellent large framed reproductions in color of modern paintings, not too well known to students (a Cezanne, a Van Gogh, a Picasso, a George Grosz, a Eugene Speicher, a Maurice Sterne, and an Alexander Brook). These were hung for at least a week at a time in six schools, without comment by teachers, allowing time for all interested students to become thoroughly familiar with the paintings. Then the art teachers in these schools asked all students, or a representative cross-section of students, to write any comments they cared to make about any or all of the paintings. The students were asked not to sign their names, but only to indicate their sex and grade in school, with the understanding that no attempt would be made to identify any comment. No directions were given except that they were not expected to write anything very profound or very clever, but to tell simply and honestly what they thought and felt about the paintings. In a few classes some of the more provocative comments were later read aloud, and more comments were collected during the ensuing discussion. About 12,000 comments were collected from about 1,000 students in grades five through twelve.

These comments were sorted until the following widely prevalent modes of response were discovered:

1. Liking or disliking the paintings
2. Liking or disliking the subject of the paintings
3. Demands for photographic realism
4. Far-fetched interpretations of what the subject represented or was doing: e.g., "The artist is trying to show how the wilderness is creeping in on the little house."
5. Fixed, dogmatic rules applied uncritically: e.g., "A portrait should always have a dull, neutral background."
6. Interpretations of the mood of the paintings: e.g., "The position of the body and the drab colors suggest sorrow and resignation."
7. A feeling of understanding, or not understanding, the

artist's intention: e.g., "I don't see what he was driving at."

8. Comments indicating special sensitivity or insensitivity to color
9. Comments indicating special sensitivity or insensitivity to design qualities other than color

A few comments on each type about each painting were selected and mimeographed. Thereafter students were asked to indicate, while looking at these same reproductions, whether they agreed, disagreed, or were "neutral" with respect to each comment. An answer sheet adapted for machine scoring was used. The directions also indicated that if a comment were true, but stupid and irrelevant, one should mark it "disagree"; and if it were neither true nor false, or partly true and partly false, or meaningless, one should mark it "neutral." The way in which the test was set up made possible two more categories of responses which were helpful in interpreting other scores:

10. Tendency to approve (to agree with favorable statements and to disagree with unfavorable statements)
11. Tendency to be "neutral" (the percentage of all statements marked "neutral")

No judgments by a jury were thus far involved except in classifying the statements as truly representing one category or another. For example, the statement "I don't know whether it is a successful portrait because I can't see enough of the subject's face" was selected by the jury as representing a demand for photographic realism. No judgment at this point was involved as to whether the comment was good or bad; only whether it was an authentic demand for photographic realism. No comments were included on which 100 per cent of the jury of artist-teachers could not agree. This was possible because there were 12,000 comments to choose from

and only 105 comments were used in the test (15 about each painting).

In selecting comments revealing special sensitivity or insensitivity to color and to design, it was necessary to decide which comments showed sensitivity and which did not. In the "interpretations of the mood of the painting," also, it was necessary to select comments which were obviously within, or very far beyond, the range of commonly acceptable interpretations. These judgments, however, were relatively easy to make, and 100 per cent agreement was secured.

Although the committee originally intended to get away from the criterion of agreement with an adult jury as much as possible, it came to feel that it would be interesting to have the jury mark the comments, and to see to what extent children of various ages approached the jury's judgment. The jury was composed of practicing artists who were also teachers—people who were presumably sensitive to art qualities and getting a great deal of enjoyment and stimulation from good painting. It was felt that if children approached the jury's way of thinking and feeling about these objects as they grew older, the chances were favorable that they were headed in the direction of greater "appreciation." The committee had become diffident about using the term "appreciation," however, so they did not apply it to the percentage of agreement with the jury. They were not sure that the jury was "right," but believed it was reasonably mature as to judgment. They therefore called this score "general maturity of response." This score is not to be taken too seriously. For example, 100 per cent agreement with the jury would probably be undesirable, since it would eliminate that individual idiosyncrasy of judgment which seems to be characteristic of people who enjoy painting. It was felt, however, that a gain from about 50 per cent agreement to 75 per cent agreement as the child grew older would probably be

desirable. Within these limits, therefore, another category of responses was created:

12. General maturity of responses (agreement with the jury)

The jury agreed almost unanimously in marking all of the statements except in the categories of "liking the paintings" and "liking the subjects of the paintings," so these categories were eliminated from consideration in arriving at the "general maturity" score. It was apparent that two equally sensitive people could look at the same painting, and both appreciate it deeply, while one liked it and the other did not. Liking paintings was essential to appreciation, but liking any given painting was not. The same reasoning would hold with even greater force with respect to the subjects of the paintings. These two categories were included chiefly to discover how they would affect other scores.

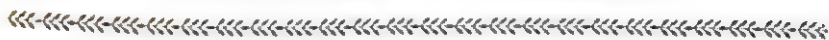
Many of these categories of responses are desirable in one period of artistic development and undesirable in another. "Demands for photographic realism," for example, would have been accepted as desirable—as making for artistic progress—in the early Renaissance, and perhaps they may still be considered desirable at certain stages of adolescence. To make scores easier to interpret, however, it was conceded that art teachers of this generation generally regard demands for photographic realism as undesirable, so this category was stated negatively in the summary sheets as "Avoids evaluating in terms of photographic realism." Thus a high score always calls attention to what most art teachers would regard as strength, and a low score to a weakness.

This test has not yet been scientifically validated, since it was developed only recently and has not yet been given to enough students to justify a statistical report on validity and reliability. Early returns, however, are very promising; at least promising enough to justify further research along these lines. The test requires some sensitivity to the mean-

ing of words, but verbal difficulties are minimized in two ways. First, students do not have to verbalize their responses for themselves, but only to indicate whether they agree or disagree with a comment which has already been phrased for them. Second, the comments are in the language of other students who have been able to put their thoughts and feelings into words, so the student is not confronted with adult concepts in adult terminology. Comments were edited only enough to remove ambiguities. Nevertheless, low scores made by students who are known to be nonverbal should be taken with a grain of salt.

Chapter V

EVALUATION OF INTERESTS



INTRODUCTION

The introduction to Chapter IV mentioned the close connection of interests with appreciations and the difficulty of distinguishing them in specific instances. Work in both areas was initiated by a Committee on Interests and Appreciations, which was later divided into sub-groups when it became apparent that techniques for evaluating interests and appreciations would be sufficiently different to justify a division of labor. The sub-committees on appreciations developed instruments, which were described in Chapter IV, to discover the ways in which students responded to literature and the arts. The sub-committees on interests developed instruments to discover and appraise interests revealed by choices of books, magazines, newspapers, radio, and motion pictures, and interests fostered by the various fields of study in school.

ANALYSIS OF THE OBJECTIVE

One of the first conclusions of the Committee on the Evaluation of Interests was that interests may be regarded both as means and as ends. When they are regarded as means, teachers try to discover activities in which pupils are already interested, and to utilize such activities in teaching pupils whatever they have to learn. They justify certain activities in the school program on the ground that they are similar or related to activities in which pupils have expressed an interest. They guide pupils who have such interests into these

activities, and direct other pupils elsewhere. They try to include in the program more activities in which pupils have manifested a lively interest. If little or no interest is expressed in a given activity, it is regarded as not likely to promote learning.

When interests are regarded as ends or objectives, however, a different approach is indicated. Teachers have to decide in what areas of activity pupils need to develop interests, and the character and direction of interests in these areas which promise most for individual happiness and the common welfare. They must then examine the evidence of interests already developed as critically as test scores in other areas of objectives, noting strengths and weaknesses, and changing the school program to build upon the strengths and remedy the weaknesses. For example, it is generally assumed that pupils should develop interests in one or more wisely selected fields of service to society, since a man who is interested in his work is usually a happier and better citizen than one who is not. If pupils, then, shortly before graduation from high school, have not developed such interests, or if their interests lie in a few fields which are inappropriate to their talents and opportunities, the school has failed in one of its obligations toward them. The character and direction of these vocational interests may also be examined. Pupils may be interested in a career primarily as an opportunity to get rich at the expense of other people, to "get to the top" against ruthless competition, and to enjoy a Hollywood conception of "success." Or they may be interested in a career primarily as a job that needs doing—as a part of a great cooperative endeavor to provide adequately for our common needs. The latter promises so much more for individual happiness and the common welfare than the former that it may be regarded as one of many criteria for judging vocational interests. In this same fashion all other areas of

desirable interests may be examined for evidence of growth in the kinds of interests which the school is trying to foster.

The Committee on the Evaluation of Interests accepted both ways of regarding interests as legitimate and necessary, but conceived its own primary function to be that of helping to evaluate interests as objectives—as outcomes rather than as starting-points of the educative process. One reason for this decision was that in the agreement with the colleges co-operating in this Study, the schools promised to provide only three types of evidence as a basis for admission to college, and one of them was “evidence of well-defined, serious interests and purposes.” Another reason was that relatively little work had been done in evaluating interests as objectives. Most of the standardized techniques as well as informal school practices attempted to discover interests as starting-points or clues in attaining other objectives; they did not evaluate the effectiveness of a school program in developing interests which were important for adolescent development and social progress.

In the course of its work, the committee had to discover and overcome three difficulties which commonly deter the evaluation of interests as objectives, and which may hamper the work of similar committees in the future. One was the unconscious assumption that little can be done about interests, that any interest is as good as any other if it is not obviously criminal, and that having no interests in important areas of activity is at most a misfortune, not a serious handicap which should be remedied by the school. The committee came to regard these assumptions as completely false. No one ever had an interest which was not *learned*, or picked up in one way or another from the environment. Even if something in the organism generates the interest, such as an interest in food, the character and direction of the interest are obviously a product of the environment. The Eskimo is

said to enjoy seal blubber and tallow candles, while we prefer beefsteak and potatoes. If all our present interests were acquired and are continually changing, new interests can also be acquired, and less promising interests can be changed for the better. A school program may be judged in part by the character, direction, and importance of the interests which it generates.

A second factor deterring the evaluation of interests as objectives among progressive teachers was the common association of evaluation with penalties and failure. It is especially obvious in this area that if pupils are given low marks or are penalized in any other way for not having interests which they ought to have, they will subsequently "fake" an interest in these areas, thus invalidating the tests without affecting their real interests. This consideration only points to the way in which almost all evaluation data should be used, but especially the data on interests. If serious deficiencies are revealed, the program should be changed to remedy them. It will do no good whatever to flunk the pupils who are deficient in these respects, nor even to criticize them. They need not even be told the judgment of the school in regard to their interests. That is primarily a matter to be discussed in faculty meetings devoted to curriculum revision, and in case conferences devoted to planning the program of individual students.

A third factor deterring the evaluation of interests as objectives was the suspicion that people who set out to implant interests in the young have in mind only adult interests. This danger was recognized and guarded against in devising instruments to discover interests which are desirable at the adolescent level. These may include some interests which would be inappropriate for adults; they may not include some interests which are indispensable for adults; and they may translate other adult interests into adolescent terms,

just as little children transform the adult interest in children into an interest in dolls. None of these considerations denies that there are areas and directions in which adolescents should develop interests. If they do not, then the school should do something to help them.

This view of interests as school objectives, which gradually evolved during the Eight-Year Study, rests upon three basic assumptions, all matters of common observation. The first is that people who have desirable interests in the major areas of life activities are obviously happier and better off than those who do not. If a man is not interested in his work, or if he is little interested in his home and family, he is so plainly miserable that the matter does not admit any philosophic uncertainty. Second, interests are the mainspring of the educational process. They practically determine what can be effectively learned. If schools, therefore, wish to develop competence in the major areas of living, they must first develop interests in those areas. Third, the common welfare depends upon the character and direction of the interests of all citizens. If these are narrow and selfish, or morbid and cruel, as in the later days of the Roman Empire, the quality of the civilization obviously declines. These three assumptions leave no choice but to find out what interests are desirable, to foster them by every means consistent with our democratic traditions, and to ascertain at regular intervals which of them are developing satisfactorily, and which of them need renewed attention.

The first principle which the committee followed in locating desirable interests was that some interests should be developed in each major area of living. These may be classified broadly as economic interests, civic interests, interests centering in the home, and recreational interests. The first three areas were sampled chiefly in the Interest Index which is described on pages 338-348, although many inferences as to

interests in these areas can be drawn from other instruments described in this report. Lack of civic interests, for example, was found to be reflected frequently in high scores on uncertainty and inconsistency on the Scale of Beliefs, and in great confusion of implications and values on the Social Problems test. These areas were also studied through many informal instruments devised for particular courses or situations and not reported here, and through standardized tests available from other sources. Vocational interests, for example, were frequently sampled by the Strong Vocational Interest Blank, by papers written in various courses, and by counseling conferences. Interests in these areas were also revealed by the instruments developed in the area broadly classified as recreational interests. An interest in books, for example, would be classified as a recreational interest, but if a student read an unusual number of rather technical books about architecture, and if in the arts and crafts (also classified as recreational) he devoted himself to drafting, to interior decoration, and to making models of houses, buildings, and communities, one might safely infer a vocational interest in architecture. Thus, all of the instruments on interests cut across the areas of activity in terms of which they are first classified.

In the area broadly classified as recreational interests, the committee distinguished five sub-areas in which interests should be developed: interests in people, in sports and games, and in the arts and crafts (including fine and industrial arts, music, dancing, drama, movies, and radio programs), in reading, and in science or scholarship—at this level, interests in the various school subjects. Interests in people were such an important element in personal and social adjustment that an instrument revealing these interests among others will be described in Chapter VI. The other “recreational” interests were sampled by the instruments now to be described.

The Reading Record

The character and direction of interests in reading which the committee regarded as most promising were the following:

1. The reading should be *abundant*.
2. The reading should be *varied* as to type and content. It should include, for example, both fiction and non-fiction; it should reflect a wide range of human experience, and deal with many subjects.
3. The reading should be *selective*, showing some concentration of interest upon subjects or types of reading suited to the reader.
4. The reading should be increasingly *mature*, gradually increasing in difficulty, complexity, and depth of insight.

It was agreed that evidence of progress in these directions could be secured through a record of reading kept by students and summarized periodically in these terms. The committee first tried out a very long and elaborate record of all reading done over a period of two weeks. This included assigned and unassigned reading in books, pamphlets, magazines, and newspapers, and asked all questions about it which any member of the committee thought would be helpful. Over 1,000 students entered their reading on this record every morning for two weeks. When the results were analyzed, it was agreed that in the future:

1. The record should involve an irreducible minimum of time and effort lest distaste for reading should be engendered;
2. The record should be filled out at stated intervals, usually once a week, in English classes. Leaving it to pupils to fill out at their convenience usually resulted in incomplete records.
3. Only voluntary reading should be recorded. Students occasionally had difficulty in distinguishing voluntary from required reading, especially when books were strongly suggested by teachers, or when supplementary reading

was required but not in specified books or amounts. Teachers were to decide what reading might be regarded as voluntary, or as indicating individual preferences.

4. The record of books read voluntarily should be kept throughout the academic year in order to get a large enough sample to provide safe inferences as to the direction of reading habits and tastes. A reliable sample of magazine and newspaper reading, however, could be obtained through a check list or questionnaire, administered annually or semi-annually.

The minimum record of books read voluntarily consisted, in most of the Thirty Schools, of notebook pages with spaces to record the author and title of each book, the date on which it was finished, and a few comments. Some teachers asked also for the number of pages in order to secure a more precise measure of "abundant" reading than the number of books. A few teachers provided a list of types of books, breaking up "fiction" (which constituted about 90 per cent of all voluntary reading) into a number of smaller categories such as school stories, adventure, mystery, love and romance, etc., and asked pupils to classify each book in terms of this list. Other teachers, who were especially interested in widening horizons through reading, asked pupils to classify each book by the nationality and period of the author, and by the period and country with which the book dealt. This was done in very broad categories. Since most of the authors read were American or English, and most of the books reflected an American or English setting, both authors and settings were classified as "American," "English," and "Other." The periods of both were classified as B.C., A.D.-1500, 1500-1800, 1800-now, and "Other" (when the period dealt with was not specified, or in the future). Most of the tallies accumulated in the spaces marked "American" and "English" from 1800 to the present, and served to remind pupils of the vast expanse of space and time which they had not yet ex-

plored in their reading. Finally, some teachers asked pupils to indicate how well they liked each book on a rough scale from 0, not at all, to 4, signifying boundless approbation.

Since most of these items could be recorded by number, referring to an item in the summary sheet, some teachers used the following form, mimeographed on notebook pages or on index cards:

Date _____ Author _____
Title _____
AUTHOR: Place _____ Time _____ Type of book _____
SETTING: Place _____ Time _____ Enjoyment _____
Comments: _____

These teachers asked pupils to keep their own summary sheet up to date as they read. This was often set up in somewhat the fashion shown on page 322.

When this sort of summary was kept by pupils, as soon as they entered a book in their reading record, they put a tally on the summary sheet opposite the type of book and under the degree of their enjoyment. As these tallies accumulated, they presented a graphic summary of the pupil's reading development in at least three of the four directions which the committee regarded as important. The total number of tallies indicated *abundance* of reading; their dispersal represented *variety* by types, periods, and places; and concentration at particular points on the first gridiron, accompanied by high ratings on "enjoyment," represented *selectivity*, which then had to be considered in terms of its appropriateness to the reader. The first gridiron also gave a rough indication of increasing *maturity* of reading, for the types of fiction listed there ranged from juvenile to adult, and the amount of non-fiction read proved also to be a crude measure of maturity, since so little of it was read by the younger pupils. In the second gridiron, almost any tallies outside the spaces reserved for American and English authors from 1800

TYPES	ENJOYMENT					
Fiction	0	1	2	3	4	T
1. Children's stories.....						
2. Animal stories.....						
3. School stories.....						
4. Sports stories.....						
5. Adventure—Western.....						
6. Sea stories.....						
7. Success stories.....						
8. Humorous stories.....						
9. Detective-mystery-horror.....						
10. Love and romance.....						
11. Historical novels.....						
12. Novels on social problems.....						
13. Tragic novels.....						
14. All other novels.....						
15. Books of short stories.....						
Fiction Totals.....						
Non-Fiction						
16. Biography, autobiography.....						
17. Books of plays.....						
18. Books of poems.....						
19. Books of essays.....						
20. Books of information.....						
21. Philosophy, religion.....						
22. Books on music and the arts.....						
23. Hobbies, practical arts.....						
24. Science, natural history.....						
25. Social problems.....						
26. History.....						
27. Travel and exploration.....						
28. Sports and games.....						
29. Vocations.....						
30. All other non-fiction.....						
Non-Fiction Totals.....						
GRAND TOTALS.....						

to the present represented a gain in maturity. These measures of maturity, however, were too crude for the purposes of English teachers who wished to measure the effects of various experimental programs, so a more refined measure was developed. This measure takes a good deal of time and some practice to use, so that it will probably be used chiefly in connection with experimental programs.

This measure of maturity was based upon a study by Jeanette H. Foster of the reading of 15,000 adults.¹ Her analysis showed that the 250 authors of fiction most frequently read could be objectively classified in six different levels of maturity in terms of the average age, education, occupational level, and general reading habits of their readers. Her placement of these authors on the various levels of maturity coincided with the judgment of the committee, looking at the list from the standpoint of the sort of maturity in reading which they wanted to develop. They therefore extended her list to include approximately 1,000 authors of fiction most frequently read by their pupils, matching each author with the authors whose maturity level had been determined objectively.²

At the same time they made a detailed classification of types of fiction and classified the works of each author in terms of this list. Authors typical of each of the six levels of maturity, from 1 (very easy reading) to 6 (very difficult reading), and of various types of fiction may be found in the following sample:

¹ Jeanette H. Foster, "An Approach to Fiction through the Characteristics of Its Readers," *Library Quarterly* (April, 1936), pp. 124-174.

² The committee responsible for the extension was composed of Harold Anderson, University of Chicago High School; Irvin C. Poley, Germantown Friends School; B. J. R. Stolper, Lincoln School; Ruth M. Ersted, Supervisor of School Libraries in Minnesota; Jennie Flexner, New York City Public Library; Jeanette Foster, Hollins College, Hollins, Virginia; and Douglas Waples, Graduate Library School, University of Chicago. Douglas Waples served as a consultant on research in reading to the Committee on the Evaluation of Reading Interests throughout its work and took major responsibility for the development of the maturity scale.

<i>Author</i>	<i>Type</i>	<i>Maturity Level</i>
Altsheler, Joseph A.	Setting	1
Austen, Jane	Character	6
Bacheller, Irving	Historical	2
Barrie, James	Character, Romance	4
Bennett, Arnold	Character	5
Boyd, James	Historical	4
Brush, Katharine	Character	3
Connolly, J. B.	Adventure	3
Conrad, Joseph	Adventure, Psychological	6
Curwood, James O.	Adventure	1
Dell, Ethel M.	Romance	1
Douglas, Lloyd	Philosophical	2

This list provided at least a standard, uniform, agreed-upon classification of fiction by type and maturity so that teachers in different schools could compare the results of their reading programs. These were summarized by teachers in a new gridiron, with types of fiction at the left and columns for the six maturity levels, unclassified, and totals for each type. Until the list became familiar, each book recorded by a pupil had to be found in the list and tallied in accordance with the type and maturity level there assigned to it. Some teachers avoided this labor by securing enough copies of the list of authors to enable each pupil to tally his own books on his summary sheet. It was feared that this expedient might lead pupils to attach undue importance to reading books at the higher levels of maturity, but when it was clearly understood that the maturity figure was largely an index of *difficulty*, and that there was no virtue in reading books that one could not understand, this fear proved to be unfounded.

The list enabled teachers to classify about 75 per cent of the fiction read by senior high school pupils. Other authors were classified by matching them with classified authors, or were tallied as "unclassified." If even 75 per cent of the fic-

tion read by a pupil were classified, this was sufficient in most cases for an individual diagnosis of the direction of reading habits and tastes in fiction.³ The list does not include enough authors commonly read by pupils in grades below the ninth to be discriminating beyond this point.

No classification of non-fiction by maturity was attempted for several reasons. It comprised only about 10 per cent of pupils' voluntary reading in most schools. It was too scattered to be easily classified. Thousands of different authors were read, but only a few by more than a handful of pupils. Frequently only parts of books were read, such as single poems, plays, essays, or chapters about a particular subject. Since so little non-fiction was read by the younger pupils, the mere number of books or of pages of non-fiction read proved to be a sufficient index of maturity for the purposes of the teachers involved. Any refinement of this simple measure would have cost more in time and effort than it was worth.

The Magazine Checklist

The record of two weeks' reading, referred to above, proved that a continuous record of magazine reading would be more burdensome than the result would justify. It also seemed to indicate that the titles of magazines read would be sufficient for purposes of evaluation, without a list of the authors and titles of stories and articles in them. While some magazines included a wide range of types of material and maturity levels, most magazines were fairly homogeneous in both respects. Furthermore, pupils read magazines rather indiscriminately, so that no safe inferences could be drawn from their choices of particular authors.

When it was decided to sample magazine reading only once or twice a year, it was found that pupils tended to for-

³ For a detailed presentation of the reading summary for one student see Wilfred Eberhart, "Evaluating the Leisure Reading of High-School Pupils," *The School Review*, XLVII (April, 1939), pp. 257-69.

get many of the magazines which they were known to have read during that period unless they were reminded by a checklist. In his Cooperative Study of Secondary School Standards, Eells had found that 108 magazines accounted for about 94 per cent of all the magazine reading done by 17,338 representative high school pupils.⁴ These magazines were listed under the following headings:

1. Popular weeklies
2. Popular monthlies
3. Picture magazines
4. "Elite" magazines
5. Non-fiction weeklies
6. Monthly reviews
7. Classroom magazines
8. Popular science
9. Sports
10. Special interests
11. Youth magazines
12. Detective, adventure, and true-story magazines
13. Motion picture and radio magazines
14. Farm magazines

Students were asked to check each magazine they had read in three columns: one indicating whether they read it seldom, occasionally, or regularly; another indicating whether they usually skimmed it, read parts of it, or read it in full; and a third indicating whether they obtained the magazine in school, at home, from a friend, a public library, a newsstand, or elsewhere. The last check had little significance for evaluation, but interested some teachers for other reasons and took almost no additional time, so that it was included for their sake. At the end of the checklist pupils were asked

⁴ Walter Crosby Eells, "What Periodicals Do School Pupils Prefer?" *Wilson Bulletin for Librarians* (December, 1937). Reprinted in *Evaluation of Secondary Schools: Supplementary Reprints*. Cooperative Study of Secondary School Standards, 744 Jackson Place, Washington, D. C.

to state what magazine they liked best, what magazines were received regularly at home, what magazines they had begun to read as a result of consideration given them in school, and what magazines they would like to have added to the school library.

The maturity level of 29 of the magazines in this checklist was determined objectively by Wert by finding the average intelligence percentile, English placement score, and score on the Cooperative Contemporary Affairs Test of readers of each magazine among 4,763 students at Ohio State University, the University of Minnesota, and five smaller colleges in the Midwest.⁵ He converted these data into an index figure for each magazine by dividing the average score of its readers on each test by the average score of readers of the *Saturday Evening Post*. The unweighted average of the three quotients thus obtained yielded an index of maturity or "quality" for each magazine, ranging from about 40 for most of the "pulp" magazines to about 200 for *The Nation* and *The New Republic*. Abundance, variety, and concentration of magazine reading were studied as in the case of books. Although it was feared in the beginning that magazine reading would not be a significant index of reading interests, since pupils would tend to read whatever magazines were received at home or in school, the variety of magazines read and its coincidence with other measures of reading development soon dispelled this fear.

Newspaper Questionnaire

In appraising students' reading of newspapers it seemed important to determine (1) what papers they read regularly or occasionally, (2) the amount of time devoted to newspaper reading, and (3) the sections of the paper which they read regularly. Since the newspapers read by students

⁵ James E. Wert, "A Technique for Determining Levels of Group Reading," *Educational Research Bulletin*, XVI, 4 (May 19, 1937), pp. 113-121, 136.

were those published in their communities, no attempt was made to prepare a checklist which sampled the titles of newspapers. Instead, a newspaper questionnaire was developed which provided spaces for the student to enter the names of the newspapers which he read and asked him to check the sections which he read regularly. Headings such as editorial, financial news, comics, book reviews, etc., were listed for him to check. The student was also asked to estimate the amount of time he spent each week in reading newspapers, and to indicate the editorial policy of each paper as "liberal," "conservative," "Republican," or "Democratic." Few students were able to do the latter accurately.

Radio and Motion Picture Checklists

The experience of the Thirty Schools indicates that a checklist is a feasible device for gathering evidence of interests revealed by choices of radio programs and motion pictures. A list of the two or three hundred motion pictures which have appeared during a three-month period may be given to students with the request that they check each picture which they have seen and indicate their degree of liking for it. In one such checklist used in the Eight-Year Study,⁶ recent motion pictures were listed alphabetically under the following headings: comedy, romance, historical, musical, sports, documentary, Western, adventure, and mystery. Including the names of the principal actors in each picture proved to be helpful in refreshing the student's memory, since titles often had little relation to the film. Students were asked to check each film which they had seen and to judge its quality. Through the use of such a checklist, data can be secured concerning (1) the number of films seen, (2) the types of films seen, and (3) the opinions of students concerning the quality of the films. In addition, the level of

⁶ The motion picture checklists used in the Eight-Year Study were prepared with the assistance of Edgar Dale, Bureau of Educational Research, Ohio State University.

quality, as judged by critics writing motion picture reviews in selected periodicals, can be determined for each film seen and a median quality level computed for the films seen by a student.

Similar checklists are useful as a measure of the extent and character of the radio listening in which students engage. One checklist used in the present study⁷ lists the popular programs heard over national networks between four and ten p.m. and all day Saturday and Sunday under such headings as variety shows, comedians, serials, religious programs, classical music, dance music, news commentators, sports broadcasts, and discussion programs. It requests the pupil to check each program which he has heard in columns indicating whether he likes it very much and listens to it whenever he can, likes it fairly well but does not go out of his way to listen to it, or dislikes and avoids it. As with the movie checklist, a tabulation of responses reveals the programs of various types listened to frequently and enjoyed most. Since both motion picture and radio checklists go out of date quickly, their usefulness depends upon their continuous revision.

The radio checklist is obviously more than a measure of interest in radio programs. For the first time in history some of the world's best music and a great deal of the world's worst music are equally available to everyone, with a perfectly free choice between them. The level of musical taste revealed by choices of radio programs is based upon a very extensive sample of voluntary behavior in a natural situation. Studies in this field indicate that high school students are at least within earshot of a radio for an average of two hours daily. They listen to the radio far more than they read. Hence, radio preferences are one of the most valid, reliable,

⁷ The radio checklists used in the Eight-Year Study were prepared with the assistance of E. Keith Tyler, Director, Evaluation of School Broadcasts, Ohio State University, and Luella Hoskins of the Radio Division of the Chicago Board of Education.

and sensitive indices now available of interests not only in music but in drama, current affairs, social problems, and the like. The radio is also unique among the instruments commonly used by schools to discover interests in that it so readily brings to light undesirable interests, or interests that are at least unpromising and a waste of time. The possibilities of this medium of evaluation have only begun to be explored.⁸

Validity and Reliability

The problem of determining the validity and reliability of activity records differs from the case of paper-and-pencil tests. A test score is regarded only as an indication of how students would respond in an actual situation calling for the ability measured by the test. It therefore has to be demonstrated that the way in which students respond to the test is the way in which they habitually respond to appropriate life situations. The test maker ideally tries to get an accurate record of how students respond to such situations and computes the correlation of their test scores with these responses. Often this is not possible, so some other indirect measure, such as marks in courses, has to be used instead, but an activity record is commonly accepted as the best criterion against which to validate a test. If the activity recorded is the objective, the only question of validity in the record of that activity is whether it is accurate. The only question of reliability is whether the record includes a large enough sample of the behavior in question to make sure that it is typical. If all the behavior relevant to a given objective were recorded, then there would be no question of reliability at all. Only when a small sample of behavior is taken do we need assurance that it fairly represents the habitual behavior of a given student.

In the case of interest in reading, the behavior which

⁸ Many promising instruments have been developed by the Radio Division of the Bureau of Educational Research, Ohio State University.

teachers were trying to develop was voluntary reading in books, magazines, and newspapers that was abundant, varied, selective, and increasingly mature. A record of such activity was secured. If the record was accurate and complete, it was a valid measure of progress toward the objective, by the very definition of validity. The behavior recorded was the objective itself—not an associated behavior which might or might not reflect the desired behavior accurately.

To find out whether the record was accurate and complete, during 1940 a member of the Evaluation Staff interviewed 51 students in the tenth, eleventh, and twelfth grades of a private, urban secondary school, who had been keeping rather extensive activity records as a part of their school program. These records included reading in books, magazines, and newspapers, attendance at plays, operas, and concerts, and choices of radio programs. The staff member explained that his interest was only in finding the facts about their records and that he had no academic connection with their school or with any college. He then talked informally with these students, asking them whether or not activities in which they had not engaged ever were recorded, and whether or not they recorded all the activities in which they engaged.

All of the 51 students interviewed said that books which they had not read were never entered in the record. In most schools in the Eight-Year Study, this was no more than prudent, for nothing was to be gained by padding the list, and the books recorded as read were discussed in conferences. Of the ten tenth-grade students interviewed, all said that all the books which they read were consistently entered. Of the 22 eleventh-grade students interviewed, ten said that not all their reading was recorded. Of the 19 twelfth-grade students interviewed, three said that not all their reading was recorded. The students who said that not all their reading was recorded explained that "trashy" books sometimes

were not entered. These "trashy" books, they said, were chiefly mystery or detective stories. Also they explained that parts of books, such as single plays, poems, essays, or stories from a collection often were not entered.

When the students were asked about the recording of motion pictures, their responses indicated that for many of them the motion picture record was quite incomplete. A few students who seldom went to motion pictures said their record was complete. However, most of the students said that not all the motion pictures which they saw were recorded. Some students said they consistently omitted recording the "poor" movies which they saw; some said they omitted recording the second feature, that is, the one they did not go to see, of a double feature program; some said that they often neglected to enter all the motion pictures which they saw, or forgot them and were unable to enter them.

All 51 of these students said that their record of plays, operas, concerts, etc., attended was complete and accurate. Such activities as attending plays and concerts, they explained, were important experiences and easily remembered; consequently all these were consistently recorded.

These interviews led to the conclusion that for these students the record of books read was accurate in what it contained but that it was incomplete. This finding would demand caution in interpreting the summaries of some students' records of books read. The quantity of reading represented in these summaries would have to be regarded as a minimum; the median maturity level of the fiction read would have to be considered in error, probably in that it would be too high. A second conclusion was that these students' difficulties in keeping a continuous record of motion pictures attended were so great as to make the use of a checklist technique a more desirable procedure. A third conclusion was that for these students a record of plays, operas, concerts, etc., attended could be kept easily and ac-

curately and represents a satisfactory method of securing evidence of participation in such activities.

Three observations need to be made. One is that under certain conditions the technique of asking students to record information about their participation in certain activities can yield valid and reliable data for the appraisal of interests. The interviews cited above revealed that for most of the students it was reasonably certain that their record of books read was both accurate and complete. Second, it must be observed that the student's attitude toward his record may be a crucial factor in determining the validity of the data. Recognizing this, the teacher should help students to understand and accept the purposes of this type of evaluation and to remove as far as possible all academic or social pressure which would tempt students to falsify their records. Third, it is important to remember that the interpretation of data derived in this fashion should attempt to take into account the conditions under which they were gathered.

The validity of the evidence secured by means of checklists is dependent upon many of the same factors as is the validity of the evidence secured by means of continuous records. A checklist requires that a student recognize, rather than recall, those activities in which he has participated; thus it demands a less difficult task of the student. A checklist, however, often must present only a sample of the many possible activities or materials and thus is dependent upon the adequacy of the sampling. The Checklist of One Hundred Magazines, for example, presents to the student only a fraction of the total number of magazines which are published. There is evidence, however, that this sample is adequate for determining the magazine reading interests of secondary school students. Students, of course, may be dishonest in responding to a checklist. Again it must be pointed out that the total situation must be considered in guarding against such dishonesty. There are no devices and no format

of a checklist which will compensate for a lack of rapport between teachers and students, for failure to prepare for the administration of such evaluation instruments, or for short-sighted use of data gathered in this fashion.

Uses of the Instruments

In making use of data gathered by means of activity records, one of the problems which teachers face is that of summarizing the data in such a fashion as to obtain a reasonably precise, yet brief, description of the interests revealed. Summaries of certain activity records for two students will be presented in order to illustrate the kinds of information about students which they make available.

Elizabeth

Elizabeth read 15 books during the year. Fiction included Mary Johnston's *To Have and To Hold*, Churchill's *The Crisis*, *The Prince and the Pauper*, Bertita Harding's *Farewell Toinette*, and *Let the Hurricane Roar*; two college stories, *Iron Duke* and *College in Crinoline*; one dog story; *The Count of Monte Cristo*; *The Girl of the Limberlost*, *Anne of Green Gables*. Non-fiction included *The Boy's Life of Will Rogers*, *Life with Mother*, *Men Are Like Street Cars*, and *Daily Except Sundays*. Eight of these books were read during the summer and seven during the school year. The class of students of which Elizabeth is a member read an average of 12 books during the summer and 24 books during the school year. She did not read books of as great difficulty and maturity as did the group as a whole. The fiction she read is distributed over Levels III (e.g., *The Crisis*), II (e.g., *Jock the Scot*), and I (e.g., *Girl of the Limberlost*); whereas the median maturity level of the fiction read by the group as a whole is IV.

In October, 1938, Elizabeth checked *New Yorker* as the only magazine she read regularly; in March, 1939, *Life*. In October, she was reading no magazine completely; in March, two—*Life* and *Look*. She was below the class median in the number of magazines read regularly and the number read completely. This evidence, together with the number of books which she read,

suggests that she does not like to read to an extent comparable with other students in her group.

Elizabeth far exceeded most of the members of her class in the number of motion pictures which she attended. She recorded seeing 39 during the summer and 86 during the school year. The median number of motion pictures attended by students of her class during the school year was 27; the range, 0 to 99. Also, she saw many of these 86 different motion pictures more than once. Evidently, then, a large amount of her leisure time was spent in viewing motion pictures. During the year, Elizabeth saw two plays: *The Boys from Syracuse* and *Abe Lincoln in Illinois*, and attended a performance of *The Mikado*. The median number of plays, operas, and concerts attended by students in her class, however, was five.

Elizabeth's five favorite radio programs in December, 1938, were Benny Goodman, Bob Crosby, Kay Kyser, Make Believe Ballroom, and Tommy Dorsey. Of the 19 programs which she checked as the ones she listened to regularly, seven were dance orchestras such as the ones listed as favorites. In addition to dance music, she listened regularly to five variety programs, three question and answer programs, two dramatic programs—Big Town and Lux Radio Theatre, and to Walter Winchell and Jimmie Fiddler. Elizabeth was approximately at the median of her class in the number of programs she heard regularly.

Claire

Claire read ten books during the summer and 35 during the school year. Five of these books read during the school year were collections of plays, such as *The Theatre Guild Anthology*; two were volumes of poetry; two were discussions of political and social problems; and four were books about journalism and the writing of short stories. The fiction she read during the school year included two volumes of short stories and such novels as *Drums Along the Mohawk*, *My Antonia*, *House of Seven Gables*, *House of Exile*, Mary Roberts Rinehart's *The Doctor*, and *Gone With the Wind*. More than half of Claire's reading was devoted to non-fiction, whereas for her class as a whole approximately 25 per cent of the titles were non-fiction. Also she read more than

the average number of books during the school year. The fiction which she read was of Levels III, IV, and V; this indicates that she was reading books of approximately the same maturity as was the group as a whole.

Claire checked eight magazines as those which she read regularly in October, 1938; and ten in March, 1939. These numbers are considerably above the group medians. In October she checked six magazines as the ones which she read completely; in March, five. Again, these numbers are above the group medians. The magazines which she read were *American Home*, *Better English*, *Life*, *New York Times Magazine*, *Reader's Digest*, *Rider and Driver*, *Quiz Digest*, and *Time*.

During the school year Claire saw 18 different motion pictures; one of these, *Grand Illusion*, she saw twice. Some of these pictures which she liked very much were *Grand Illusion*, *Four Daughters*, *Young Doctor Kildare*, *A Man to Remember*, *The Sisters*, *Brother Rat*, *Scarface*, *Gunga Din*, *Stage Coach*, *Made for Each Other*, and *Irene and Vernon Castle*. Her comments about the motion pictures which she saw and the list of pictures which she liked suggest that she chooses her motion picture entertainment with some care.

In addition to these motion pictures, Claire attended three plays, *Abe Lincoln in Illinois*, *American Landscape*, *Outward Bound*; and three musical performances, *The Boys from Syracuse*, *Ballet Russe*, and *The Hot Mikado*. This is slightly above the class median of five. Her activity record also records visits to several museums and art galleries.

In December, 1938, Claire checked eight radio programs as those which she listened to regularly. These included the Columbia Workshop, three programs of classical music, Information Please, two news commentators, and talks on politics. This number is much smaller than the median number of programs heard regularly by the group as a whole.

The leisure-time activities of these two students present two quite different pictures. One has its chief emphasis on activities such as attending motion pictures and listening to the radio with very little emphasis on reading experiences;

the other presents quite a different pattern. The one reveals interests which might be characterized as the more "popular" ones, while the other reveals interests which might be characterized as much more intellectual.

Data such as those presented in these illustrations should be of use to teachers who are concerned about the pattern of interests which students are developing. In order to use such data most effectively, it is important for the teacher to determine what kinds of interests he considers desirable for the student or the group of students, to exercise care in gathering the evidence, and to summarize this evidence in a convenient fashion. Cumulative summaries have several advantages. One is that changes which take place over a longer period of time may become evident. Another is that such summaries may be passed on from teacher to teacher as the student moves through school. Such summaries probably should not be as lengthy as the illustrations given here. However, data in tabular form similar to that suggested for books can be recorded and cumulated by students. Summary comments about the pattern of interests revealed, changes observed, and the directions in which future changes should take place might then be added by the teacher with relatively little effort.

One further suggestion about the use of such data seems warranted. Whenever possible, other evidence should be combined with the evidence supplied by such summaries in order to provide a more comprehensive description of the student's interests. The observations made by teachers both in and out of the classroom, evidence from other instruments such as the Interest Questionnaire described in this chapter, and the like, should prove useful either in corroborating hypotheses or in revealing inconsistencies which need careful study in order to arrive at a clearer understanding of the student.

THE INTEREST INDEX 8.2A

In addition to records of activities, the questionnaire has also been found useful as a method of studying students' interests. In order to investigate the possibilities of this technique, a questionnaire was developed which listed three hundred activities which students were asked to mark "Like," "Indifferent," or "Dislike." The questionnaire sampled activities which were expected to reveal interests fostered by school subjects as well as interests in certain types of relationships with other people.

Method of Selecting Items for the Questionnaire

The list of activities in the questionnaire was prepared by staff members who were concerned with evaluation instruments in the various academic fields. Each staff member examined current textbooks and analyzed classroom activities in order to identify activities which might indicate an interest developed by his field. Each activity submitted was examined critically by the entire staff to make sure that it fairly represented the interests developed by these fields and that it was actually carried on by students. All activities in which a student was apt to engage as a part or result of his work in several subjects were either eliminated or so sharpened that they became more clearly related to one field only. An attempt was also made to include items indicative of varying degrees or different depths of interest in a field: from easy and attractive activities to those involving considerable effort, hours of study, a high degree of proficiency, etc.

The items thus selected were arranged in random order in an inventory which was used experimentally in several grades in 20 of the schools participating in the Study. On the basis of the experience of staff members who interpreted the findings to the faculties of these schools and in the light

of criticisms of teachers who felt that some of the areas had not been adequately sampled or that the vocabulary of some of the items was confusing, the questionnaire was revised. This revision was also based upon an item analysis and reliability studies of the responses of 250 boys and 250 girls in typical high schools.

The Revised Form: Interest Index 8.2a

The revised form of the questionnaire consists of only 200 items and thus can be given in one study period in a junior or senior high school. The areas selected for this questionnaire are: social studies, biology, physical science, English, foreign languages, mathematics, business, home economics, industrial arts, fine arts, music, and sports. In addition to these areas, two larger categories which cut across most of them were included: reading and manipulative. These two categories are composed of items which appear in the above 12 categories and involve either reading or handwork. Thus, for instance, "To make and classify a collection of insects" is classified under biology and also under the manipulative category. The item: "To read such books as *The Life of Pasteur*, *Microbe Hunters*, *Arrowsmith*, etc." is classified under biology and also under reading. There are 16 activities in each of 11 of the above categories, 24 in social studies, 35 in reading, and 38 in manipulative. The sort of items included is indicated by the following sample. The parenthesis after each item indicates how it is classified in scoring.

1. To write stories. (English)
3. To go on trips with a class to find out about conditions such as housing, unemployment, etc., in various parts of your community. (Social Studies)
5. To visit stores, factories, offices, and other places of business to find out how their work is carried on. (Business)

6. To correspond in a foreign language with a student in another country. (Foreign Language)
7. To play baseball (either hard or soft ball). (Sports)^a
14. To learn how to cook well (in camp or at home). (Home Economics) (Manipulative)
15. To sing in a glee club, chorus, or choir. (Music)
16. To put eggs into an incubator and open one every day to see how the chick develops. (Biology) (Manipulative)
17. To sketch or paint. (Fine Arts) (Manipulative)
21. To make chemical compounds. (Physical Sciences) (Manipulative)
22. To make things of wood, metal, etc. (Industrial Arts) (Manipulative)
23. To do the arithmetic necessary in planning trips or parties for the class. (Mathematics)

Interpretation of the Questionnaire

As indicated on the data sheet on page 341, the scores give the per cent of each student's "likes" and "dislikes" in each of the categories and the per cent of his "likes" and "dislikes" for the whole questionnaire: i.e., for the 200 items. The per cent of items marked "indifferent" is not recorded but may be obtained by subtracting the sum of the "likes" and "dislikes" in each category from 100. The Data Sheet also gives the lowest and highest scores and the group median for "likes" and "dislikes" in each category.

This instrument is so simple in construction that it has been found that teachers learn to interpret it in a short time. As with most instruments, persons with greater experience may get more from it than persons with limited experience. As long as the interpreter confines himself to what he may learn about the general direction of a student's interests, the interpretation is simple and rather reliable. If, however, a person attempts to find what effect a given course offered

^a Sports were not classified as "Manipulative" because they were so nearly universal interests that they did not identify students whose interests were predominantly manipulative.

INTEREST INDEX—B.2a

DATA SHEET

	Social Studies	Biol- ogy	Phys- ical Sci.	Eng- lish	For- eign Lang.	Math- emat- ics	Busi- ness	Home Econ.	Indus- trial Arts	Fine Arts	Music	Sports	Ma- nipu- lative	Read- ing	Total
Chester															
L.....	8	0	6	44	0	6	44	56	19	25	88	69	37	8	29
D.....	58	50	69	50	81	56	25	19	19	31	0	6	34	66	39
Josephine															
L.....	21	38	56	88	88	50	81	13	25	38	94	19	21	80	52
D.....	0	0	0	0	6	0	0	0	0	0	0	50	0	0	5
Joseph															
L.....	58	81	100	75	75	81	13	13	69	31	94	56	48	84	62
D.....	13	0	0	6	6	19	56	75	6	25	6	31	18	6	21
Howard															
L.....	71	100	94	44	13	81	94	69	88	56	100	50	78	66	71
D.....	8	0	0	13	19	0	0	19	0	0	0	13	0	0	6
Boys Median															
L.....	32	40	63	36	19	42	69	31	69	31	40	62	44	40	41
D.....	15	15	6	25	42	21	9	29	4	25	27	6	20	25	21
Girls Median															
L.....	43	48	51	54	69	23	66	69	45	55	61	47	43	54	50
D.....	16	13	16	16	11	50	9	9	19	4	6	19	17	15	20

L—Like.

D—Dislike.

in the school had upon the change of interests of a group of students, certain complications arise, and rather advanced statistical treatment of the data becomes a necessary condition for arriving at valid conclusions. In the following presentation of the method of interpreting results, the relatively simple methods will be described.

Each student's scores are interpreted in relation to the group median and group range and in the light of his own scores on other categories, e.g., his own pattern of scores. The examination of scores of a student in relation to the group median and the range for each of the categories of summary will indicate in which areas the student has high or low likes or dislikes, thus establishing tentatively the deviate points in his preferences or dislikes. Thus, comparing Chester's scores with the group medians, one notices high dislikes in many areas and high likes only in three, whereas Howard has high likes in most areas and few dislikes in any of them.

One may further note the relative frequency of the significant likes and dislikes and the areas in which they occur. At this point it is helpful to examine the scores in terms of certain broad common elements in the pattern of likes and dislikes to locate the significant tendencies and characteristics of the student's pattern of interest. Thus a frequency of high likes in English, social studies, foreign language, and reading indicates high preference for verbal activities. High likes in biology, physical sciences, mathematics, and industrial arts indicate interest in activities involving things and precision manipulation. An artistic pattern is suggested by high likes in music, fine arts, industrial arts, and home economics. High likes in sports, business, industrial arts, home economics, and manipulative activities would suggest an inclination toward practical activities. If likes in one pattern are accompanied with dislikes in a contrasting one, a further reinforcement of a personal selection of activities is indi-

cated. Thus, if fairly high likes in English, social studies, foreign language, and reading are accompanied by dislikes in biology, physical sciences, and mathematics, a fairly strong case of verbal interests is indicated.

It must be noted, however, that these general patterns are nothing more than suggestions for exploring general tendencies. The areas liked and disliked group themselves in innumerable diversified ways in any individual case, and it is therefore neither possible to describe all of the possibilities, nor wise to attempt to define any one pattern precisely or to follow its implications in any one individual case slavishly.

Applying this method to the scores given above, one may note that Chester has a negative reaction to all academic activities, verbal and scientific alike. Music is the only area of high positive interest to him. In contrast, Joseph has a high interest in academic activities of all types, but shows high dislikes in such practical areas as home economics and business, and sports. Josephine's preferences run predominantly in the direction of verbal activities, with an additional interest in music and business, with no dislikes in any area but sports. Howard's interest pattern is so catholic as to arouse a suspicion of lack of discrimination.

In addition to examining the scores of a student in relation to those of other students in his group (i.e., examining them on the background of the group's scale), one must also examine these scores in terms of the *student's own scale*. Some students have high likes in many categories, others have low likes in most categories, or generally high dislikes. The total score on "likes" and "dislikes" is indicative of the general tendency of the student in terms of which his scores have to be examined. For instance, a student may be one of the highest in the group in liking music; if, however, all of his likes are high, and on his scale music is one of the lowest,

a different meaning is attributed to his score than if we consider it only with reference to the group score.

Thus in the case of Josephine, the score of 50 on disliking sports assumes great significance, because of the general absence of dislike reactions. Similarly, Chester's high dislike of mathematics, being part of a pattern of disliking all academic activities, needs to be viewed as a part of this total negative reaction, rather than as a specific reaction to mathematics. The fact that Howard's likes are uniformly high requires an investigation to see whether these are genuine interests or whether some such extraneous factor as lack of discrimination combined with a benevolent disposition is not playing a part.

One thing to be remembered in interpreting these scores is that interests are personal, and therefore a certain degree of uniqueness is both to be expected and desired. Therefore both the range and the pattern of interests should be judged in personal terms rather than by general norms. Thus, while a certain breadth of interests usually is desirable, it would be a mistake to assume that high likes in all areas indicated in the questionnaire is to be expected or is even desirable. Similarly, while negative reactions on the whole may be considered undesirable, one should expect individuals with selective interests to react negatively to some activities, while showing high positive reactions to others.

In examining group patterns, similar methods need to be applied. Thus one may note the areas in which there are tendencies toward positive or negative reactions. This can be observed by comparing the medians with the medians of other groups or by noting the frequency of high likes and high dislikes in any given area. By this method one may note the prevalence of preferences in such verbal areas as social sciences, English, and the like, or negative reactions to areas of artistic activities. There also it is important to bear in mind that a valid interpretation cannot be secured

by simply noting the areas of high likes or high dislikes. These observations must be scrutinized in terms of the total pattern as well as in terms of other data on the same group. Thus a relatively high preference for physical sciences has one meaning when this is the only area of high preference, and a different one when it is one of many. High preference for foreign language in a group with no organized experience in this field and no special aptitude in this direction usually suggests wishful thinking while the same pattern for a group with verbal ability and experience in this area can be taken to mean a thoughtful and actual interest.

Value of the Questionnaire to the Counselor or Teacher

The counselor will be interested chiefly in the configuration of the student's per cent of likes, indifferences, and dislikes in the various categories. The important point to note here is whether the picture is consistent with what is known about the student's inclinations and interests, and if some inconsistency is discovered, this lead should be investigated. When considered in connection with other information available, it should be helpful in academic or vocational guidance.

Thus the preference pattern of the student suggests the areas which can be utilized for his further development. If it seems broad enough, and sensible enough for a given student, it suggests the line of activities for him to carry on and by which he will be enriched. If an undue narrowness is indicated, the spots of positive reactions can be mobilized as a springboard for expansion of interests. Thus high interest in physical sciences would suggest that reading in that area could be used to develop interest in reading, should that be lacking. Similarly, the pattern of negative responses should suggest to teachers the areas in which remedial action may be needed or in which direct pressure should not be applied. Thus it would be futile to try to develop good work habits in English in the case of an individual with negative

responses to this area until a more positive reaction is developed. Other types of activities should be used to this end.

Since motivation is an important factor, the evidence in interests is also useful in explaining other facts about the students, such as high or low achievement in various areas, behavior in class, or activities in thinking.

The classroom teacher may be interested also in the *kinds of activities* which a given student likes or dislikes or to which he is indifferent, within particular subject-matter fields. Specific responses to individual items may be examined for this purpose and new or more subtle patterns than those revealed in the category scores may become evident. It should be noted that the emphasis in this type of examination of responses is not on the *amount* of interest which a student may have, but on the *nature* of that interest. One may find, for instance, on examining the scores that a student is at the group median in liking biology; on his own scale biology is neither particularly high nor low; but when his specific responses in this category are examined, one may find that his liking is centered on items which have to do with people, human physiology, health, etc. This knowledge should be of value to the teacher.

The classroom teacher may also make a similar use of the responses of the group. The evidence on prevailing preferences is helpful in planning classroom activities, areas to be studied or the approach to be taken. Thus exploration of printed material may be a very good way of studying a given topic for one group, while other sources must be used with groups who have a high negative reaction to verbal activities. Diagnosis of group preferences and dislikes also points to gaps in the curriculum to be filled, or unwise emphases in the present curriculum. Thus in one school an extremely high negative preference was shown for art activities. The examination of their curriculum revealed that this group had no opportunity in this field and could well profit

from it. In another case, an unusually high negative reaction to writing was traced to a large amount of required writing resulting from separate assignments by several teachers, each of whom was unaware of the total load on the students.

As in case of the individuals, the hypotheses regarding constructive action to be taken cannot be formulated validly by using the data from this questionnaire alone. These data are descriptive and as such are helpful only in suggesting hunches regarding the causes of preferences or of dislikes, yet for a remedial or constructive program it is necessary to have a fairly good idea of the cause of the interest pattern shown. Therefore it is imperative to consider these data in context of other evidence before decisions are made regarding what to do about an individual or a group.

Factors Influencing Accuracy of Results

The usefulness and accuracy of results of this instrument depend on at least two factors: the degree to which the items sample activities which are affected by the curriculum in the school in which the instrument is used, and the sincerity of the response made by the students.

The first of these may be determined by a careful examination of the specific items by the teachers who expect to use the instrument. If it is found that the items do not sample activities which reveal interests that they are trying to develop, or activities to which they would like to know their students' reactions, a similar instrument can easily be constructed which includes both.

The responses of the students will be most sincere if the instrument is not regarded as a "test" in which high scores are desirable. If the students recognize that the information which they convey through the questionnaire may be helpful in planning class work, their cooperation can be readily enlisted.

In making interpretations it should be remembered that

in this instrument the student is asked to tell how he *feels* about certain activities: whether he likes them, is indifferent to them, or dislikes them. These feelings are not necessarily an index of his performance in any of the areas sampled. A student may do poor work in class and still like many of the activities listed. Likewise a student may do very well in class and dislike many of the items. The reasons for this seeming discrepancy may be worth exploring.

For certain types of interpretations it is advisable to compute averages for boys and for girls separately, although this greatly extends the scope of the statistics which are needed. The mean, standard deviation, and coefficient of reliability of each category for the "like" scores from one sample population of 542 eleventh grade students are given in Appendix V. Reliability coefficients computed by the Kuder-Richardson formula for this sample range from .79 to .92. The median coefficient is .89, and only three categories are below .85.

A more thorough discussion of the interpretation and possible uses of this technique will be found in the next chapter. It will also be seen there that the study of interests can be used for a different purpose, namely the evaluation of personal and social adjustment. The validity of the instrument will be treated in this connection.

Chapter VI

EVALUATION OF PERSONAL AND SOCIAL ADJUSTMENT



DISCUSSION OF THE OBJECTIVE

History of the Objective

One of the concerns voiced by the schools cooperating in the Eight-Year Study was that of promoting the personal and social adjustment of their students. In an effort to clarify the meaning of these terms and to devise ways in which at least a few of the aspects of personal and social adjustment might be appraised, groups of teachers and of specialists in various pertinent fields met together. The Committee on the Study of Adolescents of the Commission on Secondary School Curriculum of the Progressive Education Association, for example, provided special help in attempting to clarify the meaning of this objective. The study of the ways in which the schools were gathering and recording evidence of students' adjustment revealed that many techniques of appraising personality and social adjustment, though they suffered from one shortcoming or another, were of promise. The work of the regional committees on anecdotal records was especially helpful in pointing to ways in which teachers might collect evidence which would give some insight into the personality problems of students.

Urged by the cooperating schools to devise more practicable means of appraising personal and social adjustment, the Evaluation Staff began an extensive study of this problem of appraisal early in 1938. Before the results of this study are presented, however, it will be necessary to attempt to dis-

tinguish between personal and social adjustment, to clarify the concepts of adjustment, and to attempt to set up a list of criteria for a method of appraisal.

Differentiation between Personal and Social Adjustment

Personal adjustment is thought of broadly as including the subjective feelings of the individual, such as feelings of adequacy and inadequacy, personal happiness and unhappiness, the adjustive reactions of the individual, the presence or absence of inner conflicting tendencies. *Social adjustment* is thought of as being directed toward the adequacy and effectiveness of a person's interaction with other people in face-to-face situations. Relationships with age-mates, older and younger people, with the opposite sex, etc., are included under this heading. It also includes the person's attitudes to the mores and standards of the group of which he is a member. It is recognized that the division between *personal* and *social* adjustment is, in some respects, an artificial one and that they should be thought of as being intimately connected and interrelated and as representing two aspects of the emotional adjustment of a person to his environment.

Discussion of "Adjustment"

There appears to be considerable difference of opinion about what constitutes adjustment. Because this term lacks clarity and may have different meanings to different persons, it is necessary to attempt to clarify the particular concept of adjustment which underlies the study to be reported in this chapter.

Broadly speaking, the investigators regard personality as a dynamic structure, which must be viewed as a whole, rather than as a collection of parts. Since personality is viewed as a product of the interaction of forces within the individual and the interaction between the individual and his surroundings, it must be seen in the light of his past history and against the background of his present environment.

The point of view underlying this investigation may be clarified somewhat by indicating how it differs from the approach which has governed certain other attempts in this field. The idea that certain behaviors, *in and by themselves*, are indicative of "good" or "poor" adjustment seems to be rather widely accepted. This point of view has been made the basis of a number of attempts to appraise students' adjustment. The procedure involves the construction of a behavior scale which lists sample statements of both "good" and "bad" behaviors. The mere counting of these behaviors is expected to give an adjustment score or index for the student.¹

Such classification of behaviors as "good" or "bad" in themselves is a relatively simple attack upon the problem. It leaves out important factors which need to be considered prior to arriving at a judgment regarding the person's adjustment or maladjustment. Two major criticisms may be made of this concept of adjustment.

It is an oversimplification which omits consideration of the individual, his motivation, surrounding temporal and environmental conditions, etc. The courts, for example, do not hold that certain acts constitute a crime everywhere and under all circumstances. Before evaluating an act, a careful study is made of the motivation of the indicted person, consideration is given to the extenuating circumstances, etc. The final judgment is also made in the light of the history of the behavior of the person. Likewise, when parents or teachers judge the behavior of children, they are aware of the necessity of attempting to determine not only *what* was done but also *why* it was done, under what circumstances the behavior occurred, and the like.

Furthermore, such a classification of behaviors as "good"

¹ For a discussion of the present status of personality measurement and of the difficulties involved, the reader is referred to Chapters I and II of *Fulcrum of Conflict*, Douglas Spencer (New York, World Book Co., Yonkers-on-Hudson, 1939).

and "bad" in themselves suffers from another oversimplification—that of not discriminating between the *condition* and the *symptom* of the condition. This may be clarified by the following analogy: an infection may be said to be a condition or a state of an organism, whereas the high fever which is apt to accompany the infection, is an outcome or symptom of the infection. Although the fever is indicative of an infection and therefore represents something undesirable, nevertheless in itself and under the circumstances it is believed to be a *desirable* adjustive reaction of the organism to the infection. In making lists of undesirable behaviors there is a tendency to use both kinds of behaviors—those which may be thought of as "conditions" as well as those which may be thought of as "symptoms"—and to neglect the fact that they are phenomena of an entirely different order and that they have to be evaluated differently.

Thus, there appear to be cogent reasons against beginning a program of appraisal of adjustment with the focus of the inquiry centering on an attempt to determine whether the adjustment of the individual is desirable or not. Determination of what specific behaviors may constitute "desirable adjustment" for a given individual is legitimate only at the end of a study of a personality, when the judgment can be based on a great many considerations. Even then it is apt to be a value judgment. Obtaining a picture revealing *how* the individual functions, what adjustive devices he employs, seems to be of greater value.

Another rather commonly accepted point of view is that adjustment consists largely in *conformity to social standards and demands*. This point of view neglects the importance of adjustment in terms of oneself, i.e., the importance of being able to handle satisfactorily one's own impulses and strivings, the importance of being consistent with oneself. It must be borne in mind that the lack of this type of adjustment expresses itself frequently in a variety of serious overt or veiled

emotional disturbances.² In this connection the following may be said regarding what must be included in thinking about adjustment. On the one hand, we have the individual with his native needs, impulses, and drives which seek satisfaction, and which undergo certain changes with age. On the other hand, we have society which has its needs and which makes certain demands on the individual. These demands on the individual vary in different cultures and depend on the age and sex of the individual, social status of the family, and similar factors. Maladjustment of the individual thus may be, broadly speaking, one of two kinds. In one instance the individual may comply to such a high degree to the demands of society that his native drives become thwarted, cramped, and distorted. In such cases the individual's behaviors with regard to society are acceptable to society, but he pays too high a price for them himself. In such an event some neurotic condition, accompanied by a good deal of anxiety and considerable personal unhappiness, may be found in him. In the second type of maladjustment the individual rebels against society, its demands and restrictions. In extreme cases such a person may suffer from society's ostracism or other types of punishment, but his difficulty, nevertheless, will be largely one of social adjustment. This is, of course, an oversimplification of the picture, yet for a broad frame of reference it is sufficiently correct. It permits us to see that in general *optimum adjustment may be thought of as a compromise between the individual and the group to which he belongs, in which each party adjusts to the other to a certain extent in order to avoid conflicts*

² The fact that educators are prone to regard as the most serious problems those of non-conformity, and to underestimate the importance of problems which are not brought to light through anti-social behavior, has been demonstrated in a number of studies. The best known of these is E. K. Wickman, "Children's Behavior and Teachers' Attitudes," The Commonwealth Fund, 1928.

within the individual or clashes between the individual and the social group.

Desirable adjustment for the individual may then be thought of as a process of maturation and adaptation during which he is able to integrate successfully (i.e., without neurotic compromises or anti-social acts) his native impulses and drives with those expectations or demands which are imposed upon him (with reference to his age, sex, social status, race, etc.) by the group to which he belongs.

The above discussion leads to the formulation of the following point of view:

1. The adjustment of the individual must be conceived as a complex of feelings and behaviors which are meaningful only when seen in relationship to each other, rather than as a series of discrete behaviors regarded as meaningful in themselves.

2. This complex of feelings and behaviors must be evaluated in terms of the status of the individual (i.e., his age, sex, position in society, etc.). The same behavior may be evaluated differently when observed in the case of a six-year-old and a sixteen-year-old, in a boy or in a girl.

3. The adjustment of the individual must be considered in terms of the relationships between his own strivings, purposes, and past conditionings, and also in terms of the relation of these to the demands or expectations of society. His adjustment must be viewed as a process rather than a state.

DISCUSSION OF THE TECHNIQUE OF APPRAISAL OF THE OBJECTIVE

Desirable Characteristics of an Instrument for Appraising Personal and Social Adjustment

Being well aware of the impossibility of evolving any single device for appraising all of the pertinent factors which need to be considered in the evaluation of the life adjustment of an individual, the staff set out to explore feasible

ways of appraising at least a few of these factors. During this process of exploration an effort was made to define the general characteristics which were felt to be desirable in an evaluation instrument for this purpose.

1. *It should be a technique applicable to a large number of students at one time.*

Since the paper-and-pencil technique is much more economical, as far as the examiner's time is concerned, than the interview, anecdotal record, etc., and thus permits testing a larger number of students at the same time, and since it rules out one of the possible subjective factors—the biases of the observer—this technique was thought to be preferable.

2. *The evidence obtained from different individuals should be comparable.*

It was felt that the form in which the data were to be collected should be such that there would be an opportunity for comparison of results. To the extent that the response-pattern of one individual can be compared with that of another or that of a group, it should be possible to discover those ways in which he is similar or dissimilar and thus gain further insight into how his personality is organized. Comparability of results might also lead to investigation of group phenomena.

3. *The technique should be indirect.*

In devising an appraisal instrument it was considered very important that the approach be relatively indirect. One difficulty which is implicit in inventories which attempt to get at the individual's private and intimate feelings is the fear and anxiety which most people experience when they feel that they are being "tested" or evaluated personally. Whereas they frequently seem able to consider certain abilities as actually extraneous to themselves and are, therefore, not

threatened when an attempt is made to measure these abilities, they usually feel defensive about obvious attempts to get at their private feelings. The anxiety aroused may be so great as to completely inhibit or invalidate the response. Thus, it was felt that the instrument should not be obviously a "Personality Test" but rather should attempt to appraise personal and social adjustment in a more indirect manner.

4. *The subject should be called on to express himself rather than to appraise himself.*

In addition to the fact that a great deal of anxiety is aroused by the demand for self-appraisal, it is also a matter of general psychological knowledge that few persons are capable of objective self-evaluation with regard to their emotions and personalities. Attempts to make a subject evaluate himself and his own emotional reactions presume a knowledge of self which is lacking in most individuals. With this consideration in mind, it was decided that asking the subject to appraise himself should be avoided; instead, he should be given an opportunity to express himself in a number of different ways.

5. *The instrument of appraisal should provide a varied response—a field upon which the subject can express himself.*

This method of appraisal differs somewhat from one of the common conceptions of a test. In many tests the subject is given a *problem* which is presumably comparable to a life situation and his performance in attaining the solution of the problem is interpreted as a measure of his ability to cope with an analogous situation in life. In an instrument which attempts to appraise personal and social adjustment, however, it was felt that it might be undesirable that the problems be thus limited by the examiner rather than revealed by the individual. It seemed that the most desirable technique to use would be that of presenting a large variety

of stimuli to which each individual might react emotionally in a variety of ways, thus providing a field, so to speak, upon which the individual might draw his own design. This means, also, that there should be opportunity for an extremely large number of configurations of response, in order that each individual might have the maximum practicable opportunity to project his personality. Single responses, then, would have meaning chiefly as they became a part of a larger pattern. Each response could be interpreted in the light of every other response. Whereas it is not possible to provide a field so large that an individual can express his whole personality, even a limited field in which the interrelationships are traceable is apt to provide a great deal of useful material.

6. *It should give the individual pattern of the personality of the subject.*

In order to get at the more detailed picture of the personality, one has to guard against the use of too broad classifications, such as "sociable" and "a-sociable." Such classifications tend to obliterate individual differences and to be useful only in very extreme cases. It was thought desirable that an appraisal instrument give a description aiming at something more than a rough categorization of the personality. This description, if it is to be useful to educators, should go beyond what is readily observable in a classroom situation. It should lead to deeper insights into the individual, his motivation, his system of subjective meanings attached to things, his values, etc. Understanding another person is an understanding of this person's acts in terms of *his* feelings and not in terms of the feelings of an outsider.

7. *It should be open to interpretation at different levels.*

It was felt that to demand from the interpreter a *certain degree* of psychological understanding is legitimate. On the other hand, it was felt that the instrument should not be so

complicated that only a person with specialized training could interpret it. Ideally such an instrument should give results which would permit deep interpretation by persons with a good deal of training and experience and still yield some useful material to persons with limited training.

Exploratory Studies

1. *Use of the Interest Questionnaire*

While the above criteria for a technique of appraisal of a personality were being considered, several exploratory studies were conducted with tests devised by the Evaluation Staff for other purposes. It was thought that since personal and social adjustment was intimately related to these other areas, a great economy would be achieved if it were found possible to draw inferences for the present objective from the results of other tests. Moreover, such an approach would be ideal from the standpoint of indirection.

Of all the tests examined from this angle, the first Interest Questionnaire, Form 8.2, gave the best results. This questionnaire provided data on the students' feeling reactions to 300 activities commonly carried on in school.³ The students responded to the items in terms of *like*, *indifferent*, *dislike*. In an exploratory study an attempt was made to discover what kinds of things and how many one might say about the personal and social adjustment of 33 college students, using the data from this questionnaire. The students selected for study were attending an institution which was known to have elaborate and detailed records on its students.

The descriptions written from the questionnaire results were compared with teachers' ratings of these students on a *Descriptive Trait Profile*,⁴ a rather flexible personality rating scale devised for the purpose of validation of this study.

³ This questionnaire has since been revised. The revised form, Interest Index 8.2a, is described in the chapter on Interests.

⁴ P.E.A. 2968 (mimeographed), University of Chicago, Chicago, Ill.

Each student was rated by four teachers. Although validation through a comparison of descriptions of personalities presents certain difficulties of a purely semantic nature, those who examined the data felt that quite similar portraits of students were presented by the teachers and by the interpreters of the questionnaire. Specifically, it was estimated that the personality sketches of 27 of the 33 students bore a remarkable similarity to the teachers' descriptions. In some cases the questionnaire revealed traits which would seem to be completely unrelated to interests as usually conceived. These results were sufficiently encouraging to justify using the interest questionnaire approach and exploring it further as a possible means of appraising personal and social adjustment.

2. *Significance of interests*

The approach taken was directly dependent upon the point of view held as to the significance of interests. This point of view differed somewhat from earlier and other current concepts of interests.

In the present study interests were approached from the point of view of the relationship between the *individual* and the *reaction* or interest. It was thought that unless we are to consider interests to be merely chance reactions, arbitrary and capricious, psychological fungi as it were, playing no part in the fundamental body of the individual's character, we must assume that they are a result of the interaction of deeper desires with environmental forces. Interest then takes on the significance of an index of emotional tendencies and of the personality pattern of the individual. It becomes the expression of the aims of the individual, conscious and expressed, or unconscious and to be inferred. Liking and disliking, accepting and rejecting activities, become significant as expressions of some of the basic elements and drives within the individual. For the purposes of this study specific

interests in themselves become rather insignificant; the emphasis is no longer on the desirability of interest within a certain field, but rather on the significance of interest for the inference of underlying urges and aims. Furthermore, interests were not thought of, in relation to this problem, as discrete, separable entities, but as interrelated and interacting.

Those who can accept this point of view about the significance of interests can readily see how an interest inventory can be used as a *projective* technique, as "a means of discovering the way in which an individual personality organizes experience, in order to disclose or at least gain insight into the individual's *private world* of meanings, significances, patterns, and feelings."⁵ The Interest Questionnaire offers to the individual the opportunity to reveal his way of organizing experience by presenting him with a large number of activities from different areas to which he reacts emotionally, in terms of *like*, *dislike*, and *indifferent*.

3. *Discussion of the Significance of Like, Indifferent, and Dislike Responses*

The exploratory study and interviews with students showed that certain inferences may be drawn from the types of responses which the student gives to the questionnaire. It was possible to do this partly on theoretical grounds, and partly because the examiners of the students' responses trained themselves to seek in the data every possible clue to the emotional state of the subjects. Thus, it was found that "like," "indifferent," and "dislike," may not be taken as meaning "just" like, indifferent, dislike, but may be thought of as having much more *affective* significance. "Like" may mean, for instance, "Is strongly attracted by it, loves." "Indifferent" may mean either no affect, or withdrawal or repression of

⁵L. K. Frank, "Projective Methods for the Study of Personality," *The Journal of Psychology*, 1939, p. 402.

affect, or an avoidance of expressing an affect. "Dislike" may express active antagonism, fear, resentment. Thus, for instance, it seemed reasonable to assume that a student who expresses a "dislike" response to a great many school activities does not "just happen" not to enjoy a large number of the listed activities but, perhaps, reveals an undercurrent of *general antagonism* to school.

DESCRIPTION OF THE QUESTIONNAIRE

The preliminary considerations and the results of the exploratory studies suggested as a next step the extension and elaboration of the interest inventory technique. This led to the construction of three inventories: Interest Index 8.2a, described in Chapter V, and Interests and Activities 8.2b and 8.2c. Each of these inventories consists of 200 items to which students respond by: *like*, *indifferent*, or *dislike*. Interest Index 8.2a consists of items relating to school studies and school subjects, whereas Interests and Activities 8.2b and 8.2c consist of items dealing with non-academic activities. It was thought that three questionnaires dealing with the intellectual, esthetic, social, and inner mental and emotional areas of functioning ought to give a rather comprehensive picture of the organization of the energies of the individual. It was further assumed that the above areas are intimately interrelated and that if attention is focussed on the interaction among them rather than on the examination of them as separable units, one ought to be able to infer a great deal regarding the functioning of the individual.

Method of Gathering Material for the Questionnaires

In order to make certain that the questionnaires contained material taken from life situations of the students, leads for the choice of the items were obtained from children. A class of junior high-school students, known rather well by one of the investigators, was told that information on children's

interests would be of use to educators, writers of radio programs, publishers of children's books, etc. They were asked how they would go about discovering such interests. After a study of the problem, the class arrived at the following methods of studying children's interests: (1) a carefully drawn up but informally administered questionnaire; (2) diary records, which were to include all activities engaged in by the members of the group, with comments as to how they had felt about them; and (3) a survey of the group as to what things its members wanted most to do or to have. The questionnaire contained such questions as: "What things do you like to do most when you are alone?" "What things do you like to do with others?" "What do you like pretending?" "What do you like to do when you feel happy?" "What do you like to do when you feel sad?" etc. The questionnaire, diary, and survey yielded a large variety of activities which formed the basis for the choice of items. As far as possible, the original phraseology of the children's statements was kept. Later a similar study was conducted in another city with a group of high school students; the resemblance between the two activity lists was striking.

Criteria for Selection of Items

In selecting items for the questionnaires, three criteria were kept in mind: (1) that the item represent a fairly characteristic or common activity of children, (2) that the activity seem to belong to one of the clusters or categories of activities which were thought to be related to personal and social adjustment, and (3) that the activity listed be not too threatening. In general, there was no effort to find single crucial items which would be diagnostic in and by themselves. Doing so would be contrary to the whole philosophy of study of personality as it has been outlined in the preceding discussion. In a sense, each item in a category may be

said to be significant only as it is viewed as a part of the total configuration of responses.

Discussion of Categories in 8.2b and 8.2c⁶

Since there seems to be no generally accepted frame of reference in terms of which a personality should be studied, the selection of categories was made in terms of the thinking of the investigators regarding some of the more important factors which need to be considered in a study of a person's adjustments. Since a possible approach toward the evaluation of adjustment was thought of as a systematic study of the individual's ways of making adjustments, rather than as an appraisal of whether or not he is "well adjusted," no categories were designed to be indicative of "good" or "poor" adjustment in and by themselves. *Each category was thought of in the light of the possible meaning it might have when examined in relation to other categories.* This must be borne in mind when examining the categories.

An effort was made to choose categories which so far as possible would yield information relative to the various kinds of adjustments the individual has to make. It should be noted that all of the information necessary for the description of an individual's adjustment cannot be obtained from the questionnaire. Information as to the environmental factors, the individual's past history, and so forth, must be obtained in some other way. The present technique aims largely at tracing some of the subjective feelings of an individual and at making inferences from these regarding the organization of his personality.

It will be seen later from the discussion of interpretation and from the sample case analysis that each student, without knowing that he is doing so, determines himself the organization of the categories by means of his reactions to the items.

⁶ The activities listed in the questionnaires are not grouped by categories; the keyed list of items can be obtained in mimeographed form from Progressive Education Association, University of Chicago, Chicago, Ill.

Depending on his responses, any of the categories may come into a dominant position in the interpretation or may come to be regarded as of minor importance in his particular case. Thus, interpretations take their lead from the student and *his* way of responding.

Nevertheless, in order to facilitate the exposition of the thinking of the investigators, in the following presentation the categories are grouped into three major areas: (1) "Organization of impulses and drives" encompasses categories which shed light predominantly on the way in which an individual handles some of his impulses; (2) "Human relationships" lists categories which are meant to tap predominantly the feelings of the student regarding social interaction of various types; (3) "Fantasy life" contains categories which are meant to reveal predominantly the extent and type of fantasies in which a student engages or which he avoids.

It should be emphasized that the above three areas are not thought of as discrete and separate entities. This classification is merely a method of organizing certain emotional dispositions which are in constant interaction. It should also be remembered that depending on the configuration, the same category may have different meanings. Furthermore, any one meaning attached to one of the categories is apt to influence the significance of some of the other categories.

1. "Organization of Impulses and Drives"

- a and b. Acceptance of Own Impulses and Severity with Oneself

Those working on the construction of this instrument felt that one of the most fundamental problems with which every growing child has to cope is the reconciliation of his primitive drives and impulses with the restrictions which social living and social mores impose on him. As has been stated earlier in the formulation of the definition of adjustment, the desirable pattern was thought of as a certain *balance* be-

tween acceptance of the primitive impulses on the one hand and, on the other hand, considerations of social expedience and actual incorporation into the individual's personality of some of the standards and restricting concepts of the social milieu. Difficulties in achieving such a balance are very common. These difficulties may be said to fall into two broad categories. The first evidences itself in a personality which continues to operate primarily on the basis of its primitive impulses and urges, and disregards or fails to incorporate the social standards and taboos. The second type of difficulty may express itself in a too rigorous repression of the impulses and their gratification and may result in a truly inhibited, extremely self-censoring and "over-restricted" personality.

Categories entitled "Acceptance of Own Impulses" and "Severity with Oneself" attempt to bring to light the student's status among his classmates with reference to the above areas of adjustment. In a sense, both of these categories aim to appraise the same area of adjustment, but approach it from two opposite poles. Thus, a very high score on "Severity" would tend to indicate that at least in certain respects the student's "Acceptance of Own Impulses" is under actual or potential censorship. A very low score on "Severity" would tend to suggest that "Acceptance of Own Impulses" functions with considerable freedom.

Examples from the category "Acceptance of Own Impulses" are: being a little sick and staying in bed all day; eating so much I can't take another bite; saying whatever comes into my head.

Examples from the category "Severity with Oneself" are: setting myself tasks to strengthen my will power; working on myself, improving myself in some way; taking a cold shower on a winter morning.

c. Preoccupation with Cleanliness

Early training in cleanliness usually represents the first demand which the social mores make upon the child to regulate his impulses. This training is often accomplished by building up strong feelings of shame or guilt about bodily functions and the body itself. Various feelings of shame and guilt, conscious or unconscious, may result in undue preoccupation with cleanliness, purity, fear of contamination, fear of germs, etc. This type of anxiety seems to be particularly common in our society. This category is designed to furnish indications as to the extent to which, and the way in which, the individual has accepted and incorporated into himself this early experience. Thus, very low likes and high dislikes in this area might indicate a lack of acceptance of these demands of society, whereas, on the other hand, very high likes and low dislikes might be symptomatic of other tensions in this area.

d. Methodical

The child's attempts to master his impulses may result in a certain rigidity of personality with a tendency to compulsive behaviors. Most of the activities in the methodical category are quite common behaviors, behaviors which are usually even encouraged by educators. They are activities which are characteristically rigidly patterned and repetitive; they also are activities which involve collecting, arranging, classifying, etc. Examples of the activities listed in this category are: copying papers to make them neat; keeping a calendar or notebook of the things I plan to do; making up catalogs and card files.

e. Aggression

Making the large number of adjustments which every child has to make, enduring frustrations, having to inhibit his impulses, invariably and quite normally produces and

contributes to the reservoir of stored hostility within the child. The expression of this hostility may take the form of overtly a-social acts; more frequently, however, it takes a more or less socially acceptable form, which serves as an outlet for the hostile feelings, without seriously imperiling the person. Categories entitled Aggression in 8.2b and in 8.2c are composed of activities through which hostile impulses frequently find an outlet. Some of these involve *overt* acts, such as: hitting someone who has annoyed me very much, always telling people the truth even when it might hurt their feelings, picking someone's argument to pieces; others involve *thinking*: thinking of what I'll do when I grow up to people who have been mean to me, looking at pictures of death and destruction.

2. "Human relationships"

f. Relationship with Family

Items dealing with activities commonly carried on in and with the family were selected for the drawing of inferences about the extent to which the student enjoys, is indifferent to, or does not enjoy his home life. An effort was made to have a wide spread of activities, ranging from such activities as having a good argument or serious discussion with the family to cleaning up after meals, washing or drying dishes.

g. Relationship with the Same Sex

This category is composed of activities in which usually only students of the same sex participate. It was thought that liking or disliking such activities as belonging to a boys' club or girls' club, staying overnight at a friend's house, etc., might be indicative of a student's feelings, particularly when reactions to these activities are seen as part of a whole set of reactions in the area of human relationships.

h. Relationship with the Opposite Sex

The items in this category were so selected that a high

score in liking them would indicate a person who attaches a value to activities requiring the participation of both sexes. This category may be broken down into:

1. Ordinary activities with the opposite sex, such as parties, dancing, etc.
2. Activities implying a stronger interest in the opposite sex than the above—making oneself attractive, courtship, etc.
3. Activities indicating a less openly displayed or perhaps vicarious interest in the opposite sex—such as reading love novels, watching others who are in love, day-dreaming about it, etc.

i. Identification with Others

The purpose of this category is to investigate the extent to which a student likes, or likes to think of himself as liking, activities which involve a strong personal interest in other people, close, intimate friendships, sympathetic taking care of others, defending the molested, etc. Many of these items are concerned with imagining things about other people, or about one's relationship with other people, rather than with actually doing things. Thus, it is possible that a student who has not yet actually established successful social relations may still like these activities. This category is designed, then, to show the extent to which the student has a value for such relationships. Characteristic items are: having a lot of close friends with whom I can talk about anything; trying to find out what a quiet shy person is really like; discussing with younger boys or girls what they like to do and how they feel about things.

j. School Activities

This category is designed to reveal the student's attitudes toward student organizations, the school, school life, etc. It is

composed of activities commonly carried on in school, such as: being an active member of a school club, being on class committees, going to school dances, etc.

k. Out-of-School Activities

This category summarizes all the activities which might reveal participation and interest in social life outside of the school situation. When considered in relation to the category school activities, it may reveal whether the student is generally sociable and enjoys all types of social situations, is generally a-sociable, or sociable in school situations but not in out-of-school situations or vice versa.

l. Solitary

This category is composed of activities in which one usually engages alone, such as keeping a diary, playing solitaire, etc. It also lists some activities which are usually sociable but are designated as solitary, such as: eating alone, going swimming, skating, bike-riding alone, etc.

m. Impressing Others

This category is composed of activities which involve pre-occupation with personal appearance, desire to be unique, outstanding, in the limelight. The following items are representative: making my handwriting unusual and decorative; having the reputation of being different or unusual; starting a fashion or a fad.

n. Leadership

Activities which involve organizing others into groups, directing groups, debating, arguing, etc., are sampled in this category. Examples of these activities are: organizing committees to plan various school affairs; being in public speaking or debating contests; etc.

o. Reactions to Authority

Activities listed in this category involve either submission to or rebellion against authority. Statements are so coded that a high score in likes is indicative of a submissive attitude toward authority, whereas a high score in dislikes is indicative of a rebellious or antagonistic attitude. Typical items are: writing papers on definite, assigned topics rather than having a free choice; being in a group where one person takes the responsibility and decides what people should or should not do.

3. "Fantasy life"

p. Birth—Life—Death

Activities in this category involve wondering about the meaning of life and death, thoughts about the origin and end of things, the meaning of eternity, and the stability and permanence of the universe. Preoccupation here might indicate the need to externalize personal anxieties and put them on a cosmic scale. Conflicts one cannot face near at hand are often projected into the cosmos, and dealt with in a philosophical way. Examples of items are: finding out how things got started; thinking about what might be the end of the world; imagining what would happen if gravity ceased to exist.

q. Fantasy

Although it is important to recognize that fantasy can play a part as an adjustment mechanism and therefore in itself is not an indication of maladjustment, it is also true that individuals who have difficulties in coping with reality may use this mechanism as a substitute for action and as an escape from actualities. This category lists a number of fantasy activities in which most youngsters engage at one time or another. Examples are: carrying on imaginary conversations

with someone I like or admire; imagining how it would feel to be rich and famous.

r. Mystery

A child who has been unduly sheltered and kept away from the realities of the life around him may develop not only distorted notions regarding his environment, but also great curiosity and preoccupation with "the secrets of adults" and other mysteries. The items in this category attempt to sample the different "mystery-interests" of children and adolescents. Such statements as the following are found in the questionnaire: having people "forget themselves" and talk freely; listening to other people's phone conversations.

s. Magic

Every child, at least in part because of his relative incompetence as compared with adults, in his efforts to deal with his environment tends to resort to magical means, such as good luck charms, avoidance of symbols of bad luck, etc. Great dependence upon these symbols may reveal a feeling of incompetence and a need to resort to "superior powers" for help. This category lists some of the activities which involve using magic, such as: carrying a good luck charm; making up little games or schemes which will bring luck if they come out right; seeing if a hoped for thing comes true if I concentrate on it.

t. Dramatics

This category is composed of theater arts activities—those involving writing and production of plays, and those involving taking specific roles. It can be interpreted both as revealing interest or lack of interest in the theater arts *per se*, and it can also be interpreted as revealing the wish or fantasy life of the individual. In this connection an examination of the types of roles which are preferred is particularly

interesting. Examples of items are: thinking up plots for plays; taking the part of a wicked or dangerous person in a play.

u. Humor

This category is composed of activities which have to do with the appreciation or expression of humor. Humor may be thought of as a way of relieving tension. It also is frequently an accepted, subtle way of expressing hostility. This is particularly clear in playing practical jokes and other such forms of humor. The items in this category also serve to make the whole questionnaire lighter in tone and more entertaining. Examples are: drawing cartoons; seeing plays which are "take-offs" on dignified people or institutions; reading or writing funny poems or limericks.

INTERPRETATION OF THE RESPONSES TO THE QUESTIONNAIRES

The questionnaires are scored in terms of the per cent of the items in each category to which the student responds with *like*, and the per cent to which he responds with *dislike*. The per cent of *indifferent* responses may be readily calculated by subtracting the sum of the above two scores from 100. As will be seen presently, the interpretations may be made on two levels. For a quick overview of the student's interests and adjustive trends, one may examine his tabulated per cent scores in the various categories on the Summary Sheet. This takes little time and gives a fair but rather general picture. A much more detailed study of the student may be made from the examination of his specific responses to individual items in the questionnaires.

Interpretation of the Scores on the Summary Sheet

A student's score on a category acquires meaning in two ways: when viewed in reference to the group median, and

when viewed in relation to the student's scores on other categories.

No scores are considered high or low *per se*. The student's scores are always examined in the light of the scores of other students in the *group* in which he is living and working. It is possible, however, to single out categories in which he ranks high or low in his group in likes or dislikes. From examining these categories it is possible to draw certain inferences about the student. For instance, it frequently happens that a student has low likes on the academic interest questionnaire (8.2a), but has high likes on all the sociable categories in the non-academic interest questionnaires, or vice versa. A student's likes and dislikes may group themselves not only in this broad manner but may also group themselves in greater specificity. One may find, for instance, a student who is high in likes in categories involving precision in work, such as physical science, mathematics, industrial arts, and methodical, whereas he may be low in likes in categories involving greater freedom of action and self-expression, such as fine arts and dramatics. Again a student might be low in liking such sociable activities as are listed in the categories same sex, opposite sex, sociable activities in school, and sociable activities out of school, and at the same time be high in liking fantasy, mystery, magic, etc. Many different configurations are thus possible.

The final picture is derived from the way in which the individual student reacts to a great many fields of activity: academic interests, sociable activities, and activities which indicate his attitude toward himself. In linking these seemingly quite different fields, the interpreter attempts to discover the common elements which make the student's response to academic situations understandable in terms of the way in which his personality is organized. The fact that the meaning of a given score in a category may change with the response of the student to other categories is an important

consideration. For instance, if a student responds to leadership by liking 80 per cent of the items and also comes out high on fantasy, but comes out low on most of the categories dealing with sociable activities, one is justified in raising the question as to whether or not the high liking of leadership indicates wishful thinking. Careful study of results thus far has indicated that if one watches for the inner consistency of the picture presented by a student, one learns to discover facts about his fantasy life and learns to single out his wishful responses. The fact that with some students the questionnaires are apt to reflect their wishes rather than represent their actual behaviors is an important one and should not be regarded as something which makes this technique invalid. On the contrary, is not this gaining of insight into the inner mental life of the child the most difficult but important part of the problem?

It frequently happens that a student's category scores are *generally* high or low in likes, indifference, or dislikes; i.e., one finds students who are "high likers," "low likers," "highly indifferent"—high or low, that is, in relation to the group medians. When there is a general tendency to respond in a certain way, deviations from this tendency become important, even though the deviations may not be apparent at first. If, for instance, a student is below the group median in likes in all categories, but near the median in some categories, and at the same time is one of the lowest in the class in his scores in likes on other categories it becomes evident that *his scale* has a smaller area, but that there still is a differentiation in his response.

In each case it is necessary to examine all three scores, like, indifferent, and dislike. A student may have an equal like score on two categories, but the fact that he feels differently about the activities in each may be evidenced by a strong dissimilarity in his dislike scores.

Generally, the process of interpreting the summary sheet

is as follows: The interpreter first picks out the highest likes in relation to the student's other scores, and attempts to seek common elements in these categories. The same is done for the high dislikes and high indifferences. This examination includes also a consideration of the categories which the student likes or dislikes least.

Interpretation of Responses to Individual Items

Although this approach to personality study attempts to procure quantitative data on emotional tendencies and dispositions and seems to do so rather successfully, for a deeper understanding of a student a more detailed analysis is necessary. This is done by an examination of his responses to individual items and is a procedure which is particularly important for gaining an understanding of the dynamics of the student's behavior. Here again the same main principle of interpretation as is used with the category scores is applied. First the likes, then the dislikes, and then the indifferences for individual items in each category are taken and each time an attempt is made to single out the common elements which characterize or run through the given group of activities. This examination frequently reveals *new* categorizations peculiar to the individual whose responses are being examined. For instance, two students may have very similar scores in the total number of likes on the category opposite sex; one may like only the items concerned with actual sociable activities; the other, however, may like only those items showing a vicarious interest, those involving fantasizing, reading romantic novels, etc., and be indifferent to or dislike the actual sociable activities. In the same manner, one may observe that a student may consistently like or dislike all the items involving speaking before a group, regardless of whether the activity appears in a foreign language class, mathematics, or in a social situation. Many such individual categorizations have been traced.

With regard to the study of specific responses within a category, it must be borne in mind that the meaning of any given response to any item must be again examined in a twofold way. First it must be examined from the point of view of the particular pattern which it reveals for the student; i.e., from the point of view of the types of activities within a given category or in different categories that the student likes, dislikes, and is indifferent to. Second, these specific responses must be examined against the background of the responses of the same age and sex group. To make such a comparison possible the staff is preparing a table of responses of students to every item in the questionnaires. These tables are based on a study of responses of a large number of students and will show how the boys and girls of different school grades have distributed their likes, indifferences, and dislikes. Thus, for the evaluation of the meaning of a specific response it is important to know that less than 10 per cent of both boys and girls of all grades from seven to twelve mark "dislike" the item: "Talking in halls and locker rooms." The significance of a student's specific responses obviously changes depending on how the majority of his age mates respond to the item.

Discussion of Total Scores on 8.2a, b, and c⁷

We note that whereas the bulk of Lyle's interests on the academic interest questionnaire (8.2a) are in the upper quarter, in the non-academic questionnaires (8.2b and c) none of his likes on any of the categories is high enough to place him in the upper quarter of his class. On all except one category in 8.2a he shows zero dislikes. The only category in which he has any dislikes is sports. It begins to look as if sports were differentiated from the other interests on 8.2a by Lyle, possibly because this area of activity involves dealing

⁷ This description was made from the material obtained from the questionnaires above. The teachers' descriptions were made and held by them until the completion of the interpretation.

SAMPLE ANALYSIS OF RESPONSES OF ONE STUDENT

TABLE I

Scores of One Student and Medians of his Class on Three Interest Questionnaires

Lyle O., Age 12 years, 6 months
Mid-Western Private School
7th Grade, Class of 22 boys

Case No. 13

Category	Likes			Dislikes			Indifferent	
	Rank in Group	Per Cent Scores	Median	Rank in Group	Per Cent Scores	Median	Per Cent Scores	Median
<i>8.2a, 20 Boys</i>								
<i>Upper Quarter</i>								
Fine Arts.....	1	88	38	17-20	0	25		
Music.....	1	75	14	18-20	0	58		
Manipulative.....	2	76	50	17-20	0	20		
Industrial Arts.....	2-5	94	75	18-20	0	4		
Mathematics.....	3	75	20	18-20	0	38		
Business.....	3	69	31	15-20	0	23		
Total (a).....	3	63	36	20	1	28	36	36
English.....	3-4	63	26	19-20	0	42		
Foreign Languages.....	3-4	63	22	18-20	0	37		
Reading.....	3-4	54	33	20	0	22		
Physical Science.....	4	88	48	17-20	0	14		
<i>2nd Quarter</i>								
Biology.....	6	63	30	16-20	0	23		
Social Studies.....	6	46	23	18-20	0	24		
<i>3rd Quarter</i>								
Home Economics.....	12-14	19	26	18-20	0	26	81	48
<i>Lower Quarter</i>								
Sports.....	16-17	31	50	9-11	19	17	50	33
<i>8.2b, 20 Boys</i>								
<i>8.2c, 22 Boys</i>								
<i>2nd Quarter</i>								
Fantasy.....	6-7	63	52	12-15	0	19	37	29
Humor.....	6-8	66	58	10-11	14	12		
Magic.....	9-11	25	25	13-15	19	27	56	48
Family.....	9-12	49	49	17	6	22	45	29
Authority.....	10	22	17	18-19	22	52	56	31
Dramatics.....	10-12	38	38	14-17	6	25	56	37
<i>3rd Quarter</i>								
Methodical.....	11-12	36	37	21	8	35	56	28
Severity.....	13-15	24	30	15	32	38	44	32
Acceptance of Own Impulses.....	14-15	34	37	12-13	26	31	40	32
Impressing Others.....	14-17	19	31	19-20	19	38	62	31
Opposite Sex.....	15-16	22	31	18	12	24	66	45
Total (b).....	15-17	30	38	18-19	13	24	57	38
<i>Lower Quarter</i>								
Birth-Life-Death.....	16	18	32	15	20	30	62	38
School Activities.....	16-18	26	39	18	12	16	62	45
Aggression (b).....	17	18	48	4	50	22		
Mystery.....	17-18	31	56	13-15	0	19	69	25
Aggression (c).....	17-18	14	42	6-8	36	27	50	31
Leadership.....	17-19	13	60	5-7	19	13	58	27
Total (c).....	18	28	39	21	21	31	51	30
Identification with Others.....	18	18	32	9-12	25	25	57	43
Out-of-School.....	19-20	8	44	15	0	15	92	41
Same Sex.....	20	0	43	9-16	10	10	90	47
Solitary.....	20-21	22	38	9-12	28	27	50	35
Preoccupation with Cleanliness.....	22	8	34	19	16	38	76	28

Figures falling in the upper quarter in the indifferent column are italicized.

with other people and dealing with certain environmental realities. (Incidentally, having zero dislikes for any of the academic activities is most unusual.)

Whereas Lyle is above the median in indifference in practically every category on 8.2b and c, he is above the median in indifference in only two categories on 8.2a—home economics and sports. This seems to point to some very important differentiations in the organization of his energies. Probably his indifferences in 8.2a should be examined separately as they may be equivalent to dislikes in his case. Evidently, for some reason which we do not know, Lyle has a value for the “academic”—and either accepts or feels he should accept everything which seems to fall into this classification.

On 8.2b and 8.2c, it is interesting to note that on *his* scale of interests fantasy is highest, whereas all of the categories involving interaction with other people (with the sole exception of family) fall below the median. It would seem, again, that he distinguishes in some way between activities with other people and the things that go on in his mind.

Lyle has only three high dislikes on 8.2b and 8.2c: aggression (b), and aggression (c) and leadership. We see in this a strong avoidance of asserting himself, openly, with other people. Furthermore his high indifference in the category authority, coupled with the very low dislike and what is, on his scale, a fairly high like of it, make us feel that he is a boy who has accepted a certain set of adult standards and avoids expressing any criticism or questioning of it. In a sense he seems to be a boy who is pretty thoroughly subjugated by the world of adults. It is startling to note that Lyle has zero likes in the category dealing with activities with the same sex, and has only 8 per cent likes in the category dealing with sociable activities out-of-school. This is very unusual for a seventh grade boy or any boy for that matter. Actually, he shows on these questionnaires a slightly higher interest in

the opposite sex than in the same sex. Usually this is reversed among seventh grade boys. However, his interest in the opposite sex is not high enough so that it could be called an outlet for his sociable feelings. It would rather seem that he does not avoid it to the extent that he does the same sex. (An examination of Lyle's specific reactions in this category reveals that he likes only three items. These items are only remotely connected with this category and do *not* involve any activities with the opposite sex—they deal rather with *learning facts* and with *daydreaming*.)

Lyle's low likes in the category solitary seem contradictory to the picture we have been getting of him. In his case, however, we tend to think that this low score is an indication of a tendency in him to avoid admitting to himself (or to others) that he does not have a normal play-life with other boys and girls. If this hunch is correct then we may say that Lyle may, in himself, have a value for or feel a lack of satisfaction in the sociable area, but that the full realization of the fact that he misses something in life is too painful for him and he attempts to convince himself that he is really indifferent to it.

Discussion of Reactions to Specific Items on 8.2a

Since Lyle has dislikes only in the category sports a detailed examination of his responses in this category may be fruitful. Such an examination reveals that he dislikes: to play baseball, to play basketball, and to do setting-up exercises. The strength of these dislikes is particularly impressive when we recall that they are the *only* items which he so marked on the whole questionnaire. We notice further that he is indifferent to all the team games. He likes only such highly individualized sports as: to play horseshoes, to shoot with bow and arrow, to play golf, etc.

In social studies we notice that Lyle is indifferent to all "social action" items, such as: taking part in a campaign

against countries or business firms which treat people unjustly; attending public meetings to protest against something you regard as unfair; getting people to vote for certain candidates, etc. On the other hand, he likes those items which deal with study, reading, and history. His interest in social studies seems to be largely an academic one.

Discussion of Responses to Individual Items on 8.2b and 8.2c

We may take first categories on which Lyle expresses high dislikes (for him).

Leadership. In this category Lyle is indifferent to almost all the items except that he likes to speak at a club or class meeting, and likes organizing a hobby club. Both of these are explainable in terms of his interest in academic pursuits. He dislikes: organizing groups to vote in a certain way in school elections, organizing a protest meeting in or out of school (cf., social studies), and being captain of an athletic team. This latter item is disliked by only two other boys in the whole class.

Aggression. In this category Lyle dislikes such aggression as: throwing spit balls, throwing things when I am mad, playing a joke on a teacher (disliked by only three other boys), picking a fight when I am in the mood, and telling someone what I think of him. He has only five likes out of a total of 33 items in this category, and these likes are distinguished by the fact that again they are not open expressions; in fact, they seem to represent what may be called fantasizing about his aggressions. He likes: thinking of what I'll do when I grow up to people who have been mean to me, checking up on things that teachers say in order to find out if they are true or not, reading about real crimes and how criminals get caught, and thinking about how to become the cleverest, richest, hardest financial genius in the world.

Authority. The striking thing here is that Lyle is very indifferent to authority. His very indifference seems to indicate

a certain submissiveness. We notice that he likes: having a teacher lead and supervise a free-time activity, having a teacher outline in detail what should be studied and how to go about it, and being on a committee where the chairman makes the decisions instead of allowing a lot of discussion (he is the only boy in the group who likes this item). We draw from this the inference that Lyle is happiest in a teacher-controlled situation, and that for some reason or other pupil-controlled or pupil-dominated situations contain some sort of threat to him.

The avoidance of asserting himself in leadership and aggression and his apparent liking of following adult authority and avoidance of interaction with other youngsters, makes us think that the hostility which he must have toward his group must be expressed through isolation from the group rather than through open conflict, except perhaps in a very spotty and spasmodic way. This isolation from the group is probably expressed in his fantasy activities and also by using his intellectual interests as a way of achieving superiority (in his own mind) over other youngsters. We consider that he has adopted too early the adult-approved pattern, without having gone through the necessary stages of really arriving at it. This, we tend to believe, has fixated him on an emotionally immature level of development. It is interesting to note that he likes: having people take me for older than I am, discussing things with older people, etc. The world of adults seems to threaten him much less than the world of other youngsters.

This interest in older people is in striking contrast to his seeming lack of warm, intimate, friendly interest in his own age-group. We notice for instance, that Lyle is the only one in his group who dislikes: trying to find out what a quiet, shy person is really like, standing up against a group and defending a person who has been picked on, etc. Such responses make us think that he is probably essentially very

shy himself. We tend to feel that while his constellation of academic interests may seem "mature," there is a great dependence upon adults. Thus he seems to fear those situations in which he is unprotected. We notice, for instance, that he dislikes: talking to strangers, taking a long trip all alone, having my parents go off on a long trip, etc. This again seems to point to that odd combination of adultish and infantile qualities in Lyle upon which we have remarked before.

In connection with this we note that Lyle is in every instance indifferent to items which are concerned with personal appearance. There are only two categories to which he is more indifferent than he is to preoccupation with cleanliness—out-of-school activities, and same sex.

Some general comment should be made about the possible meaning of Lyle's indifferences. We are inclined to interpret them in two ways: in part, they seem to represent a withdrawal of his energies from the sociable areas and throwing them into the academic area; in part, they may be an escape or protection from the reality situation. The very great indifference (over 60 per cent) in such categories as same sex, out-of-school activities, school activities, and opposite sex is really very striking. We do not interpret this as meaning that Lyle does not have or never had any desire for social interaction, but rather we interpret it as meaning that for some reason, and in some way, he finds such interaction difficult and disturbing. We tend to think that he would like to be able to get along with other people. He likes, for instance: carrying on imaginary conversations with someone whom I like or admire, imagining situations in which I might be a hero, planning long adventurous journeys, etc. (In connection with this we note that he does *not* like the reality versions of these statements—i.e., he does not like: trying to describe my innermost feelings to a friend, standing up for someone who has been picked on, taking a long trip all alone, etc.) Thus we see an important discrepancy between his

fantasy life and his attitude toward his real life. We also notice a tendency to project a great many of these wishes or unsatisfied desires into the future—he likes for instance: planning my future family, daydreaming about the future, listening to fantastic plays about the future, and, on the other hand, imagining what I would do if I could live my life over again.

In conclusion, one may say that Lyle probably does not get into open clashes with adults and is very likely to be academically a good student. His age-mates may elect him to class offices, but probably few of them, if any, accept him as a real member of the group. A number of youngsters are apt to be annoyed by him and make him the butt of their jokes. Lyle's main difficulties seem to be that although or because he has accepted prematurely the standards and values of a certain group of adults—his own emotional development has been warped and arrested.

*Statements checked and written in by teachers who filled out the
Descriptive Trait Profile*

Shy, retiring, academic minded boy. Likes science especially. Retreats from all social functions. Adult in thinking and associations. Brother so much older. Father and mother very brilliant. Lyle suffers from asthma and many allergies and heart weakness. Fear of death is strong.

Observable propelling drives? For perfection and truth in scientific approach. Strong questioning mind—extremely modest—introvert.

Vital, active, efficient, well-organized and concentrated in his attack on school work.

In thinking through a problem tries within the range of his ability to obtain a wide range of facts and considers and weighs them impartially before arriving at a conclusion.

Outstanding interests: Science—impersonal scientific research. *Anything* but people.

Thought of as being only moderately boyish in dress, activities and interests, and physique.

Average looking. Timid soul type. Not physically strong. Pleasant boy, however.

Too secure with parents—and himself—not enough with boys his age—adultish in standards.

Holds rigid standards for himself—very self-critical.

Follower—and yet respected because he knows his stuff.

Tendency toward daydreaming, fantasy—Lyle is an introvert—but in the scientific sense.

Ordinarily contented, satisfied, serene. Tends to make the best of situations even when they are unpleasant.

Calm, composed, even, level-headed, well-balanced. Expresses his emotions freely and is not either uncontrolled or over-restrained.

Generally flexible and adaptable; adjusts readily to new situations, to changes in routine, etc.

Self-confident in a calm way, estimates self fairly correctly, accepts own assets and liabilities fairly realistically; is not over-modest nor has the need to brag.

Is fairly well-poised.

Shies away from students of the same sex.

Is respected though not a prominent member of the group. His friendship is sought and he enjoys popularity and attracts students.

May not have any strong individual attachments, yet responds in a moderately friendly and interested way to the opposite sex.

RELIABILITY

The reliability of each category of scores on the two questionnaires was computed by the Kuder-Richardson formula for a sample population of 1,000 students, divided evenly between boys and girls and among grades seven to twelve in several representative schools. The results, along with the

range of scores, the mean, and the standard deviation on likes and dislikes in each category, are given in Tables in Appendix VI. In general, the coefficients of reliability range from .53 to .86, the median coefficient for likes being .78 and for dislikes .75. Only three categories of likes and six of dislikes have a reliability coefficient lower than .70. While a higher degree of reliability would be desirable, considering the intrinsic variability of behavior in this area, the reliability of other tests in this field, and the way in which one score is continually checked against another, the obtained reliabilities were considered sufficiently high for the purposes of these tests and for the manner in which they were interpreted.

VALIDITY

The problem of validity of a technique of appraisal is one of paramount importance. It is a complex problem, however. On the long road at the start of which are the assumptions which underlie the technique and at the end of which are the final interpretations or descriptions of a subject, there are many points at which validity should be questioned and scrutinized. As it has been stated above, the degree of effectiveness of the present method of study of personality was checked upon at the very beginning of the study when 33 college students were described and these descriptions compared with the school records of these students. Similar informal studies have been conducted as work progressed. These studies helped in guiding the staff in its experimentation with untried methods and suggested the abandonment of certain ones which were not found fruitful. The following is a presentation of some of the findings on validity to date.

Discussion of the Evidence on the Validity of the Questionnaires

Broadly speaking, validity may be broken down into two parts: (a) validity of the instrument as such, and (b) validity of the interpretation of the results.

Genuineness of response. One important element involved in the validity of any instrument of appraisal is the so-called genuineness of the response of the subject. By genuineness of response, in this instance, is meant the extent to which the response represents the real feelings of the individual. If, as may be the case in such an instrument, the response represents wishful thinking, it is nevertheless genuine, for the wishful thinking is an important part of the individual's feelings. It is possible to have genuineness of response without making valid interpretations of these responses, although it is difficult to see how the contrary might be true.

One would naturally expect some fluctuation in category scores from year to year because of growth factors. If the response were not genuine one would expect marked and unpredictable fluctuations in category scores from year to year. One would be dealing with chance or random reactions. If, however, after having made allowance for the growth factor, there still is a fairly high relationship between the category scores one year, and the scores on a retest a year later, one might be justified in concluding that there is constancy, and therefore genuineness of response. The following table shows the results obtained when correlations were run between the category scores of 48 boys and 56 girls who responded to the questionnaires in the seventh, eighth, or ninth grades one year, and in the eighth, ninth, or tenth the next.

These data seem to indicate that having made allowances for the growth factor there is still a high degree of consistency of response, and therefore of predictability. It would seem justifiable to assume that genuineness of response was a contributor to this constancy factor.

In preparing the questionnaires it was felt important to learn how students feel about this approach. In an attempt to determine this, toward the end of the third questionnaire was placed the item: "Answering questionnaires like this."

TABLE 2

Product-Moment Correlations of Scores Obtained One Year Apart

Category	48 Boys	56 Girls
Leadership.....	.78	.48
Fantasy.....	.77	.77
Life-Death.....	.76	.77
Identification with Others.....	.74	.70
Aggression (c).....	.72	.65
Total (c).....	.70	.81
Self-acceptance.....	.70	.70
Total (b).....	.68	.75
Humor.....	.68	.74
Cleanliness.....	.68	.49
Mystery.....	.66	.60
Methodical.....	.65	.59
Out-of-School.....	.62	.57
Aggression (b).....	.61	.61
Dramatics.....	.58	.69
Non-Identification.....	.58	.67
Magic.....	.58	.70
Severity.....	.57	.63
Family.....	.55	.68
Opposite Sex.....	.53	.64
Authority.....	.46	.20
Same Sex.....	.40	.78
Solitary.....	.44	.55
School Activities.....	.34	.68

It may be seen from the following tabulation of responses to this item that girls in all grades enjoy the questionnaires more than the boys, that students in the lower grades like them more than the older students, that in most grades more students marked this item *like* than *dislike* and that only in the case of the tenth grade boys did as many as 41 per cent of them mark this item *dislike*.

Discussion of the Evidence of Validity of Interpretations

1. *Validation through information from the school.* During the course of the present study the questionnaires were ad-

TABLE 3

Per Cent of Students Responding Like, Indifferent, and Dislike to "Answering Questionnaires Like This"

Grade	Number		Per Cent of Boys Responding			Per Cent of Girls Responding		
	Boys	Girls	L	I	D	L	I	D
7	78	91	42	32	26	66	26	8
8	60	50	47	30	23	78	8	14
9	164	177	41	30	29	57	28	15
10	97	176	32	27	41	49	28	23
11	114	200	42	24	34	43	23	34
12	126	95	30	32	38	35	30	35

ministered widely in a number of schools and in several of these schools the Evaluation Staff agreed to furnish written descriptions of some of the students' personalities in order to check on the correctness of the interpretations derived from the questionnaires. The faculties in the schools selected the students for this study before the questionnaires had been administered. The only information on these selected students which the staff had was the name, age, grade, and sex of the student and the responses to the questionnaires; on the basis of this information a rather detailed description of the personality of each student was prepared.⁸

While the written descriptions of the students were being prepared by members of the Evaluation Staff, teachers who knew these students best rated them on the Descriptive Trait Profile. The Profiles were held by the school until the school received the interpreters' descriptions of students. As an additional check on the descriptions derived from the questionnaires, the teachers who had rated these students were asked

⁸ The case which was presented in the preceding section is one of these studies. It was selected for presentation because it was shorter than most.

to read these descriptions carefully and to make marginal notes, especially in instances when they disagreed with the picture presented. By this method 16 case studies were made.

2. *Method of appraising the extent of agreement and disagreement with the material submitted by the schools.* Since the present approach to personality study is thought of essentially as a technique which aims to bring out some of the outstanding features of a personality, different patterns of organization of energies of individuals, it was felt that the final validation should employ methods suitable for such material. This made it impossible to attempt to arrive at some single index or coefficient which would represent the degree of validity of the interpretations. It was thought further that the problem of validation of descriptions of students derived from the interest questionnaires involves the examination of the cases from three angles. First, there must be an appraisal of the *comprehensiveness* of the description of the students, the extent to which the analysis brings out a number of significant facts about the student (significant from the point of view of the counselor and classroom teacher). Second, there must be an appraisal of the *degree of consistency* or inconsistency between the interpretation of the questionnaire results and the material presented by the school on the same students. Third, since the descriptions derived from the questionnaires at times attempted to go beyond what the classroom teacher might know about the student, a judgment had to be made regarding the reasonableness or probability that these *inferences* were valid in the light of all the information available on the student. The same judgment had to be made in cases when there was an actual disagreement between the two descriptions; the teacher's judgment could not be accepted as necessarily infallible any more than could that of the interpreters.

Because none of the simpler statistical methods could be used to measure the degree to which two pictures of a per-

sonality coincide or differ, or to determine which picture is more likely to be psychologically correct, it was thought that the opinions of a number of competent judges would form the best evaluation of this study. In other words, the criterion of enlightened common sense seemed to be the most feasible method of appraising the validity of the interpretation.

Sixteen judges were selected and they were asked to guide themselves by the following questions in making their judgments: (1) Would most reasonably competent people tend to agree or disagree that the *same* tendency or characteristic of the student was commented upon by the interpreters and by the teachers, even though they may have described this characteristic in different words and in a different context? (2) From my experience with children and adults, from my observations of human behavior and motivation and from *all* facts presented in this case, which of the two statements about the student seems more likely to be correct—the one made by the interpreters or the one submitted by the school?

The judges were asked to use the following procedure in making their evaluation of this material:

1. Read through the interpretations of the interest questionnaires carefully.
2. Read the comments of the teachers, marginal and otherwise, including the information from the Descriptive Trait Profile.
3. Make a statement regarding (a) the degree of *comprehensiveness* of the picture of the student, (b) the degree of *agreement* between the interpretation and the data from the school, and finally, (c) in cases of disagreement, a judgment regarding which of the two pictures seems most reasonable or *valid* in the light of all the information gathered on the student.

A list of statements was prepared for each of the three questions (a, b, and c) on which judgment was sought. The judges were instructed to check the appropriate statement but to regard these statements as merely suggestive and to

feel free to make their own statements. The tabulation of statements checked or written in by the judges will be found on the following pages for all 16 cases. Since each case was judged by four judges the total number of judgments for each of the three questions should normally be $16 \times 4 = 64$. Because some judges checked more than one statement, the actual number of statements is often above 64.

LIST OF JUDGES

- Peter Blos, Institute for the Study of Personality Development, New York City.
- J. F. Brown, Professor of Psychology, University of Kansas, Lawrence, Kansas.
- P. S. de Q. Cabot, Director, Cambridge-Somerville Youth Study, Cambridge, Massachusetts.
- Frank S. Freeman, Professor of Education, Cornell University, Ithaca, New York.
- Robert J. Havighurst, Professor of Education and Secretary of the Committee on Human Development, The University of Chicago, Chicago, Illinois.
- Josephine R. Hilgard, M.D., Fellow in Psychiatry, Institute for Juvenile Research, Chicago, Illinois.
- L. L. Jarvie, Director of Guidance and Curriculum, Rochester Athenaeum and Mechanics Institute, Rochester, New York.
- Harold E. Jones, Director, Institute of Child Welfare, University of California, Berkeley, California.
- Jean W. Macfarlane, Director of Child Guidance Study, Institute of Child Welfare, University of California, Berkeley, California.
- George J. Mohr, M.D., Clinical Staff, The Institute of Psychoanalysis, Chicago, Illinois; Associate Professor of Criminology, University of Illinois Medical School, Urbana, Illinois.
- Willard C. Olson, Director of Research in Child Development; Professor of Education, University of Michigan, Ann Arbor, Michigan.
- Daniel Prescott, Professor of Education, The University of Chicago, Chicago, Illinois.
- Fritz Redl, Professor of Psychology, Wayne University, Detroit,

Michigan; Division on Child Development and Teacher Personnel, Commission on Teacher Education, The University of Chicago, Chicago, Illinois.

Helen Ross, Research Associate, The Institute for Psychoanalysis, Chicago, Illinois.

Verner M. Sims, Professor of Psychology, University of Alabama.

Herbert R. Stolz, M.D., Assistant Superintendent in Charge of Individual Guidance, Oakland Public Schools, Oakland, California.

TABLE 4

Judgment as to the comprehensiveness of picture of student, the usefulness of this information to the counselor or teacher.

<i>Statement</i>	<i>No. of times checked</i>
1. The description of the personality of the student is very clear and comprehensive; it should be of real value to a counselor.	15
2. The analysis seems to have come very close to several of the central difficulties of the youngster; it should be of help to the counselor.	29
3. Although the interest questionnaire did not obtain a consistent and clear-cut picture of the student, the study unearthed some important hypotheses about him.	12
4. The description from the interest questionnaires is too vague and equivocal to make a judgment.	2
5. The statements in the interpretation could apply to anyone—there is nothing which seems to apply to this youngster specifically and alone.	1
6. Many dominant characteristics mentioned by the school are missed completely in the interpretation.	8
<i>Total</i>	67

TABLE 5

Judgment as to the degree of agreement between interpretation and data from school.

<i>Statement</i>	<i>No. of times checked</i>
1. The picture presented is highly consistent with the material submitted by the school.	17
2. There is agreement on important aspects of personality, disagreement on the less important.	9
3. There is general agreement between the report of school and the interpretation, but the interpretation seems to over-emphasize or exaggerate certain aspects.	8
4. There is agreement in part, but there is a lack of verification by the school on details.	7
5. There is excellent agreement in some parts, whereas in other parts there is marked disagreement.	8
6. The school gives a "surface" picture of behavior, whereas questionnaire results describe "central" or "underlying" behaviors. This makes a comparison difficult.	1
7. There is marked disagreement in most areas; only in minor points is there agreement.	2
8. There is little agreement between the interpreters' analysis of the major outline of personality and the version presented by the school.	8
9. Neither of the reports gives a clear-cut picture; therefore, a comparison is difficult.	1
10. Insufficient data from school for making a judgment.	2
11. There seems to be no relationship between the interpretation and the description presented by the school.	1
<i>Total</i>	64

TABLE 6

Judgment as to which picture seems most reasonable, or valid in the light of all the information gathered on the student. (In cases of disagreement, or in cases in which the interpretation goes beyond the material presented by the school.)

<i>Statement</i>	<i>No. of times checked</i>
1. The interpretations which go <i>beyond</i> the material submitted by the school are psychologically very consistent with the total picture.	19
2. The description derived from the interest questionnaires seems more convincing. I tend to accept it as being more likely to be psychologically correct.	18
3. Even though the school's description of the youngster's behavior and the interpretation of his feelings as revealed through the questionnaires do not seem to coincide, it is very probable that each is valid at its own level.	9
4. Analysis seems to have hit upon the central themes of conflict, a fact which renders it especially valuable for the counselor.	2
5. The questionnaire results help to get at some of the causes of the picture of maladjustment painted by the teachers.	1
6. The conclusions of the analysis give perspective and psychological meaning to teachers' statements.	1
7. The questionnaire results and the school report supplement each other, though I regard the questionnaire as the more valuable psychologically.	1
8. The questionnaire interpretation is more penetrating than the school material. The school	

description, while helpful, has a few inconsistencies; and it is more of a surface description.	1
9. The description presented by the school and the description derived from the interest questionnaires supplement each other to form a consistent picture of the student.	19
10. There are too many contradictory statements from the school to make a judgment.	1
11. Insufficient data from school to make a judgment.	1
12. There are too many contradictions in the material to make a judgment.	2
13. The description presented by the school seems more convincing or plausible. I tend to accept it as being more likely to be psychologically correct.	9
<i>Total</i>	84

These three tables indicate a preponderance of opinion in favor of the inferences about students drawn from the questionnaires. Of 194 judgments which may be classified as favorable or unfavorable, 157 favor the questionnaires, while 37 express some criticism or indicate a preference for the materials presented by the school. Of the latter, 31 express only the following criticisms: many dominant characteristics mentioned by the school are missed in the interpretation (7), the interpretation seems to over-emphasize or exaggerate certain aspects (6), there is little agreement between the interpretation and the school's version (8), and the school's description seems more plausible (9). Some of these were not intended as criticisms for they were frequently expressed by judges who preferred the version given by the interpretation. When it is recalled that the material presented by the school was the result of several years of close association with and study of students, while the interpretation was based on three short tests by investigators who had

never seen these students and knew nothing else about them, the preponderance of critical opinion in favor of the questionnaires is encouraging.

POSSIBLE USES OF THE QUESTIONNAIRES

It may be well to indicate at this point that paper and pencil interest questionnaires do not necessarily constitute the best method of studying interests. It is possible that skillfully conducted interviews, direct observation, etc., may yield much richer, more dependable material. On the other hand, it may be that one of the advantages of a questionnaire is the fact that a mass of comparable data are secured on a large number of students at one time. This material can be used for studies of individuals or for studies of groups or for studies of shifts of interests occurring with age in boys and girls.

Value to the Counselor

1. It is expected that persons who work out a few of the individual interpretations and who begin to see the intimate relationship between the so-called "academic" interests and the emotional dispositions of the individual, will begin to view the in-school behavior of youngsters quite differently.

2. The questionnaires afford the opportunity to look at a student from a *new* angle—the expression of his likes and dislikes in a great many areas. These one examines in terms of the individual and in terms of how he compares with the other members of the group.

3. The questionnaire results suggest a number of hypotheses about the student—point to directions which ought to be investigated. The questionnaires are expected to serve the function of a time-saving device since they point out specific areas which have to be investigated first. Such investigations are not blind trial-and-error searches for information, since they are based on an hypothesis and since the

area investigated is naturally connected with some aspect of the student which is of importance to the educator.

4. On the basis of the information derived from the picture of the interests and on the basis of the information obtained from other sources, it is expected that courses of action will suggest themselves. These remedial steps will be based on a knowledge of the student's abilities, on a knowledge of his academic interests, and on some facts regarding his personal and social adjustment.

The question of the extent to which it is legitimate to discuss with students their scores is being asked repeatedly. Some teachers even feel that a description of a student derived from the questionnaires should be read to the youngster. Those who have worked with the questionnaires take a very definite stand on this point. It is felt very strongly about 8.2b and 8.2c that the scores should *never* be shown to a youngster, just as the youngster is never shown his Intelligence Quotient. There are two main reasons for taking this stand.

In the first place, by making the students self-conscious about the questionnaire, by revealing to them the nature of the categories on which they expressed themselves, one would spoil the chances for administering the questionnaires again. The next time the answers would be apt to be much less spontaneous; the student would tend either to give the teacher what he thinks the teacher wants him to give, or give whatever ideas he has regarding his liking for a given category as such. It would be very similar to giving the students the key to questionnaires and asking them to respond to items as they are arranged under the various categories instead of having the statements in a random order. This consideration applies to 8.2a as well as to 8.2b and 8.2c.

The second reason which makes letting the students see their own scores seem undesirable is the injury which this may do to them. When one constantly sees adults who take

numerical scores, medians, etc., as if they were absolute and infallible realities, one can easily imagine the damage which may be done to a youngster who would suddenly be confronted by the fact that he scored way below the median of the class in liking his family or that he came out highest in the class in disliking it. Even if the scores were absolutely correct representations of youngsters' feelings, pointing them out to the student would not alter these feelings, but would be apt to increase the self-consciousness and, therefore, the conflict about these feelings. There seems to be a very common misconception in the minds of many people that the mere pointing out of a fact to a person has therapeutic effects. This misconception may be due to two things. In the first place, it is true that in relatively simple matters, pointing out a fact to a person often makes this person watch himself in this respect or makes him actually change his behavior. For instance, when a student consistently misspells a word or has difficulty in constructing a sentence, pointing out his shortcoming to him may have beneficial effects. In the area of feelings or emotions, however, the pointing out of a tension or conflict or the pointing out of a symptom of a tension often tends to aggravate the situation.

In the second place, this misconception may be due to an incorrect understanding of the word "insight," which is frequently found in psychological literature. Contrary to the popular notion, an effective guidance worker, psychologist, or psychiatrist does not *give* insight to his client, but, when this is indicated, so works with the client that the latter *gains* insight into himself. *Giving* insight, instead of allowing the person to develop insight, often only strengthens the block which prevents the person from understanding what is really operating in him. To help a student *gain* insight requires a great deal of skill and considerable experience. The classroom teacher who may have some qualms about an undertaking of this sort, will nevertheless be able to gain cer-

tain insights which will assist him or her in manipulating the environment of the youngster as a means of making it easier for the student to make the necessary adjustments.

It is somewhat less dangerous to let students see their scores on 8.2a. In certain situations this may be permissible, much depending on the type of youngster one is dealing with and much depending on the relationship between the student and the interpreter of the questionnaire. One should be always cognizant of the fact, however, that such a discussion is almost certain to make it impossible to give the same questionnaire again. Moreover, the student is apt to take his score, as compared with the median of the class, as evidence of a permanent characteristic of himself, perhaps as evidence of an inherent lack of interest in the subject, perhaps even as evidence of his inability to do well in this area. Trying to correct this by telling a student: "Now just snap out of it, John, you can be interested in this as much as anyone else!" can hardly be expected to stimulate a real interest.

In cases of students who are really eager to learn more about themselves and their performances on the questionnaire, it is suggested that, without showing them their actual scores and the median of the class, one could pick out the highest interests of the individual, mentioning to him that they seemed to be *his* highest interests and pointing the discussion in the direction of what *this* student actually enjoys doing, what he actually enjoys at school, etc. The areas of low interests, as revealed by the scores, do not have to be discussed with reference to the questionnaire but may come up for discussion naturally, as the outcome of the whole conversation. The above approach in which one starts with the area of outgoing feelings and interests of the student is thought to be much more positive. This positive approach is apt to make the whole discussion a pleasant and spontaneous one and is apt to cement the relationship between the counselor and counselee rather than create a breach.

The Administration of the Questionnaires

Questions relative to the method of administration have been brought up by a number of teachers. Some seem to feel that the situation under which the questionnaires are administered has a great deal to do with the results.

It is thought best to present the questionnaires rather casually, perhaps as part of a survey of the school or as part of a study of pupils' interests. Certainly the validity of the results is considerably reduced if one tells the students that the school wants to find out "everything about their personalities" or if one singles out a troublesome student and lets him take the questionnaires by himself or under the immediate supervision of some stern adult. Preferably the questionnaires should not be given at a time when they draw the students from an activity which they particularly enjoy. Their resentment will probably reflect itself in their responses. The traditional "test" situation should be avoided as much as possible and every effort should be made to make it a pleasurable experience.

The fact that most of the items in the questionnaires were furnished by youngsters indicates that frank statements *can* be obtained from them. The fact that such responses can be obtained only by a person in whom the children have complete confidence, because of this person's tact in dealing with their feelings, must also be borne in mind.

SUMMARY

In concluding this chapter it may be well to point out some of the main features of the present technique of study of personal and social adjustment. These features may be summarized as follows:

1. *Indirection.* It is felt that the questionnaires do not appear to the students to be obviously a "personality test," and that therefore they do not arouse the anxieties which many such tests evoke. They have been found to be actually en-

joyed by a great many children. Most of the items in the questionnaires have been obtained from children's diary records of their daily activities. Whenever possible, youngsters' language was preserved in the inventories.

2. *Flexibility.* The inventories do not attempt to discover whether the student does or does not fall into one of a group of patterns prearranged by the investigator. Rather they attempt to provide a field upon which, with certain limitations, the student may trace his own pattern or profile. The subjects are thought to reveal their various affective trends through the configuration and the interrelation of their responses.

3. *Aims at a dynamic instead of a static picture.* This method attempts to reveal how a student operates or functions, what adjustive devices he employs, how he feels about various activities. This aspect of the method is expected to be of particular practical usefulness.

4. *Aims at gaining insight into students' motivation.* Insofar as it is possible through the examination of specific responses to discern common elements in new groupings of likes and dislikes, one is frequently able to see what lies behind these feelings. This gives useful clues as to how to motivate the student's interest in some other activities.

5. *Tends to make a student's academic likes and dislikes understandable in terms of the organization of his personality.* It is felt that only too frequently there is a dichotomy in our concept of a personality. The thinking life of a student is thought of as a discrete, separate unit determined by his I.Q. and "special abilities" and unrelated to his needs, drives, and goals. The approach outlined above aims to bring to light certain common trends in the individual which evidence themselves both through his academic interests and other activities. Should it be possible to give a classroom teacher an instrument which will enable her to relate the strivings and the goals of a student and the possible satisfac-

tion of these goals to work on certain academic problems, the opportunity to make education meaningful to children would be increased greatly.

6. *Final results are descriptive rather than definitive.* Instead of having the final picture a score or series of scores, it is a brief personality sketch or study. This sketch is derived from the way in which the individual student reacts to a great many fields of activity: academic interests, sociable activities, and activities which indicate his attitudes toward himself.

7. *Questionnaire results are inferential.* The present approach should not be thought of as a "test" or as an instrument which is meant to give conclusive evidence regarding a student's personality. The results are inferential. The interpretations should always be regarded as hypotheses which, when combined with other information on the student, might prove useful to the counselor.

Chapter VII

INTERPRETATION AND USES OF EVALUATION DATA



The preceding chapters have explained the development of evaluation instruments in several major areas of objectives. References to methods of interpretation and uses of these instruments were confined to single instruments or pairs of instruments. Other problems of interpretation and uses were encountered when a whole program of evaluation was developing. The present chapter is devoted to these problems.

Methods of interpretation and uses of evaluation data were determined largely by two factors. One was the conception of the functions which interpretation was to serve; the other was the character of the data and the assumptions on which they were based.

Functions of Interpretation

Since the main purpose of evaluation was to help teachers improve their curriculum and guidance, the first function of interpretation was to *translate* the evidence from columns of figures into descriptions of behavior which were intelligible and useful to teachers for this purpose. Such translation occurred on three levels: single scores or bits of evidence, whole instruments, and batteries of instruments.

At the first level, even a single score on a test usually carried no self-evident meaning. What, for example, did a score of 11 per cent on crude errors in the test on interpretation of data mean? It seemed to be low (desirable); it was actually high (undesirable) as such scores went; but in a group which had had little training in this ability, it might be below

the median, and better than was to be expected from this student. Thus each score had to be translated, at least in the mind of the interpreter, in terms of the behavior which it represented.

Each score, however, was only a part of the larger pattern of behavior revealed by a given instrument. At the second level of translation, therefore, each score had to be interpreted in the light of the other scores on the same instrument, in order to see the larger tendencies in behavior in this area and their dependence on one another.

This process was continued with scores from a battery of instruments at the third level of translation. Thus, scores indicating inability to get accurate meaning from quantitative data, combined with evidence of general ability in logical discrimination and skill in quantitative techniques, might indicate that the difficulty lay only in failure to devote the necessary attention and persistence.

This level of translation made possible the second function of interpretation: to suggest hypotheses regarding the possible causes of the strengths or weaknesses of individuals and groups. To locate such causes, it was necessary to consider not only all available evidence of present status but also the history of development up to this point, and the relevant factors in experience in and out of school. This was entirely possible when the data accumulated gradually, and when teachers had known their students for a long time.

Finally, it was the function of interpretation to suggest hypotheses regarding constructive measures to remedy the situation. This was a step requiring thoughtful judgment, not a decision that could be made automatically. Usually it was necessary to consider the objectives of the school, the pattern of goals of the individual, as well as the demands made on him by life or school activities in order to decide which shortcomings needed to be remedied. A wise judgment regarding the methods of remedy required, in addition, insight into

human behavior and the methods by which that behavior could be controlled and changed.

The Nature of the Data and the Assumptions Underlying Them

The process of evaluation was composed of two elements which on the surface seemed contradictory, and which traditionally had been held to be contradictory. In the first place, any form of appraisal is essentially an analytic process. To see each individual clearly and accurately and to observe the differences among individuals more precisely, it was necessary to break up larger complexes of behavior into their component parts and to get as accurate measures of each as possible.

Thus, in the course of the Eight-Year Study, reference was often made to "breaking up" objectives. Separate instruments were constructed to appraise each area of objectives, and in many cases each aspect of specific objectives. This type of approach could easily be identified with "atomism," that is, with an assumption that human behavior is composed of isolated reactions, each of which can be understood, explained and appraised as a separate entity.

However, evaluation in the Eight-Year Study has also adhered to the second, synthesizing function of appraisal. One of the most influential psychological principles guiding the work has been the assumption that the essential characteristic of human behavior is its organic unity, and that various aspects of it function in close relationship with each other. It was clear that no single aspect of human behavior would be understood without reference to the total pattern of behavior. Similarly, it was clear that usually no single type of growth could be fully achieved without some progress in all others. While an uneven development was expected toward certain objectives, such as thinking, attitudes, interests, social adjustment, and so on, no one aspect should be developed too far without some growth in other important aspects of de-

velopment taking place at the same time. Thus, if logical thinking were cultivated without much attention to emotional and social maturation, not only would the development of thinking be handicapped; personality maladjustments might also appear as a result of too uneven a rhythm of growth. Similarly, the possibility of rational and objective social attitudes was greatly limited unless a certain degree of maturation took place in social interests.

This basic assumption found expression at several points in the development of the evaluation program. One of these was the conception underlying the comprehensive set of objectives. The areas of objectives described in the first chapter were not chosen arbitrarily or accidentally. In formulating objectives and in classifying them, an effort was made to include such a range of the significant aspects of human growth that, taken together as goals of development, the areas of objectives would represent a unified and related development of the whole person. Thus the term "comprehensive" used in conjunction with objectives referred primarily to the range of aspects of human growth viewed as an organic unit.

The idea of relatedness of behavior was also expressed in the structure of the instruments developed as well as in planning the series of instruments. Thus, each instrument attempted to diagnose a pattern of closely related behavior aspects rather than isolated behaviors. For example, in developing the test to measure the ability to apply social values to controversial problems, an analysis was made of the behaviors involved in this process. The ability to see implications of social values broadly or comprehensively was considered to be one of them. At the same time, it was evident that some people, while seeing issues broadly, also indulged in inconsistent and irrelevant reasoning. While their scores on comprehensiveness might be quantitatively the same, the meanings of these scores differed depending on what logical

qualities were shown at the same time. Further, the question of the nature of their values entered. A broad and comprehensive awareness of values and their implications might involve a consistent or inconsistent, homogeneous or ambivalent pattern of those values. This pattern might be what is commonly called "democratic," or "undemocratic." Recognizing the relationship of these three types of reactions, namely comprehensiveness, logic, and values, it was necessary to construct a test permitting the diagnosis of each of these behaviors in a context involving the others. The test provided for each type of reaction and permitted a description of them in their relationship to each other.

While each instrument was constructed to appraise specific behavior related to specific objectives, the relationship of these behaviors to the total behavior pattern of an individual was not forgotten. In many cases instruments were frankly devised as "mates" to each other, because it was clear that the behaviors measured by them were strongly influenced by each other, or because it was recognized that certain kinds of behavior needed to be checked in different content. Thus the instruments measuring general social beliefs were supplemented with others appraising the application of these beliefs in concrete situations and the logical thinking involved in such a process. The evaluation of free reading was conducted hand in hand with the evaluation of responses made to that reading. Information and application of information were found to be importantly related and some instruments appraised both with reference to the same content. Similarly, recognition of the strong relationship between interests and thinking made it necessary to secure evidence on interests in all areas in which logical thinking was appraised, so as to be able to diagnose weaknesses in thinking in relation to interests in the same areas.

Often an effort was made to secure supplementary evidence from a series of instruments on certain characteristics

appraised directly in one instrument. Thus the tendency to go beyond data or to be overcautious was directly measured in the test on interpretation of data. Supplementary evidence on the same tendencies could be gained from other tests also. For this reason some scores were retained even though their statistical reliability as separate scores was low, for the reliability of the conclusions increased as the same tendency was shown in many different instruments.

Thus, in a sense, the series of major instruments composed a related *battery*. Each instrument was a part of a comprehensive plan for evaluation, designed to correspond to related behaviors within a unified pattern of development. Thus the synthesizing function of evaluation was expressed in the structure of the instruments as well as in the relationship of the instruments to one another.

As a result, what the interpreter found was not a series of isolated data, but a series of data which fitted into a pattern of behavior relationship. His job was facilitated because the required synthesis was not to be brought about from a planless series of isolated bits of evidence. Certain generalized relationships were inherent in the very nature of the data. His task was to detect the variations of individual and group patterns within this general framework.

Illustrative Case Study

To illustrate the problems encountered and methods of reasoning and inference fruitful in synthesizing a range of data, a case study is presented on p. 409. An effort was made to use the types of data actually securable in a public school and to analyze them as they were analyzed by the school staff. A deviation from the school's procedure was necessary, however, in the order of presentation.

Usually a case study of test data is made when a decision is necessary regarding some problem of an individual or group. The occasion may be that of choosing a program of

studies, a difficulty observed by some teacher, a behavior problem requiring explanation, or some inconsistency observed in the data themselves. The nature of the problem usually determines at which point the analysis of information begins and what sequence the consideration assumes.

The case of Jane came to the attention of counselors and teachers when they surveyed the data from a battery of instruments prepared by the Evaluation Staff and found that the impressions of Jane secured from these data differed from the ones prevailing among the school staff. For this reason the investigation proceeded first to locate some of the outstanding conflicting impressions and then to examine data relevant to explaining them. However, the data are here presented not in the order in which they were secured or analyzed in the school, but in the order of their explanatory value for the subsequent data.

Background Data

Jane is a senior in a large public high school and has come to it through a junior high school on the same campus. Several teachers have thus known her for some time. She is considered an average, normal child, so much so that, according to the counselor, she has scarcely been noticed. She has never created any trouble, has done her work fairly well and, except for occasional difficulty with her Latin teacher, has behaved as a "good" student. Her I.Q. is 120 (Terman group) which is in the middle of the range of her group.

Standardized Achievement Test Scores

Her percentile scores on standardized achievement tests taken over the preceding two years were as follows:

Year I		Year II	
Algebra.....	55	English Usage.....	87
English.....	84	Spelling.....	64
French.....	85	Vocabulary.....	98
Latin.....	100	Literary Comprehension....	92
Medieval History.....	99	Reading Rate.....	85
		Literary Acquaintance.....	98

Apparently Jane has a high level of achievement in the usual subject matter skills and information. With the exception of algebra and spelling, her scores are at or above the 84th percentiles.

Two questions suggest themselves at this point. First, one notices that her scores on mathematics and spelling are conspicuously lower than the others and one wonders what may be the cause for that. Secondly, one is curious about how Jane's standing in the class on achievement scores compares with her abilities as measured by intelligence test scores.

Examination of the range of scores for the group revealed that Jane tends to stand higher on achievement tests than on the intelligence test scores. One notices also that the areas of her high achievement are areas of high verbal content which suggests a special proficiency with words and possibly difficulties with areas and processes requiring the use of other techniques and symbols.

Teacher Reports

A look at the teachers' reports to her parents reveals the following:

Algebra—Teacher has little to say, except that Jane has difficulty with learning mathematics, especially when it comes to application of quantitative concepts to practical problems.

English—In general, Jane understands what she reads. Some of the modern poetry presents difficulty. She needs to increase her speed of reading. As far as free reading is concerned, she shows "appreciation, acquaintance, and scope in her reading." Her literary background is satisfactory, especially with reference to literary criticism. When in a hurry, Jane makes unreasonable mistakes in spelling. "Jane knows better." Organization of materials is excellent and presentation acceptable. Excellent work habits.

French—Reads with comprehension, speed, and accuracy.

Has good memory for words. Understands and remembers grammatical principles. Reads smoothly and knows rules of pronunciation. Responds orally in fluent speech. Written work could show improvement in application. Is much interested in foreign people and their contribution to civilization. Does individual research work in music for her own pleasure. Work habits excellent, though lack of preparation was evident in the last two tests during the two weeks preceding the report. Has intellectual interests in Romance languages and their development. Is studying Spanish in her leisure time and corresponds with a foreigner in that language.

Latin—Has keen power to get thought from foreign language without translation. Vocabulary is very good; grammar and pronunciation good. In applying fundamentals, written work is better than oral work. For the past six weeks has made no effort to do more in silent reading than the minimum requirement. Is unique on occasion in applying historic-cultural materials, but frequently fails to come through. Work habits are bad. Does not pretend to do things on time. Intellectual interests sometimes very high, sometimes very low.

Social Studies—Good mastery of such skills as reading, map work, use of graphs and charts, library books. Knows a satisfactory number of historical facts. Reads more than average, though mostly nature books. Work habits are steady and persistent. Has intellectual interests in cultures different from her own.

A few things stand out in these reports. First, with the exception of reports from the teacher of Latin, teachers' reports are consistent with the results of the standardized tests. The mathematics teacher reports difficulty with algebra, and the English teacher comments on Jane's "unreasonable" spelling. One wonders whether the teachers' reports were

based on or influenced by the achievement tests, but the reports were written before the tests were given. The fact that the Latin teacher reports difficulty with Jane's work habits, while her achievement score in Latin is very high, suggests several possibilities. First, the Latin classes may emphasize objectives not measured by the achievement test. The Latin teacher may have been unduly influenced by Jane's slump during the last six weeks, and may be applying pressure to get her out of it. Jane may also have had some special difficulty with the teacher which may have influenced the teacher's observations. Finally, Jane's proficiency with words may have caused her to be bored by the class work, which she mastered all too easily. Each one of these points can be checked easily enough in the school situation. According to the counselor, the Latin teacher was the only one who insisted that Jane develop a modicum of precision and care with details. Others seem to have been satisfied with more general accomplishments.

Behavior Descriptions by Teachers¹

The descriptions by teachers of several of Jane's behavior traits are rather diverse and on the whole non-conclusive. On the 15 traits there described, usually the teachers of French, social studies, and occasionally English, place her higher on any given trait than do the teachers of mathematics and Latin, particularly the latter. Thus, in assessing her imagination, the French teacher describes her as "generally imaginative," the social studies teacher as "specifically imaginative," mathematics teacher as "imitative" and Latin teacher as "unimaginative." Similarly, according to the French teacher, she is highly analytical, but according to the mathematics and Latin teachers, limited in her power of analysis. In most of the 15 characteristics, she gets the

¹ The forms developed by the committee headed by Mr. Eugene R. Smith were used. These forms are described in Part II of this volume.

highest as well as the lowest ratings. This suggests several possibilities. First, the teachers may have had insufficient opportunity to observe Jane on all characteristics, and therefore may have given somewhat invalid reports. The teachers may also have rated Jane according to her achievement in the class, thus being influenced by what is called a "halo effect." It may also be that Jane's difficulties in academic achievement influenced her personal relations with each of the teachers concerned and hence affected her actual behavior in class.

Summary of Counselors' Interviews over Two Years

Due to the loss of her parents, Jane lives some distance in the country with her grandmother and aunt. She has consequently had little companionship with other children and is thrown a great deal with older people. Moreover, the grandmother and the aunt do not get along well, and Jane feels that she often has to take the brunt of their differences with each other. Jane feels that her ideas are "foreign" to those of her grandmother and aunt, and she suppresses them at home, "for the sake of peace." When the difficulty with her work habits in Latin was pointed out to her, Jane said she was in the habit of leaving work to the last minute and rushing through with it, a habit indulged in by many "bright students." Since she got good grades, "why bother?" As to her difficulty with Latin, she felt that she could get more out of the language by herself.

Concerning her personal life, Jane confesses that she cannot work with other people, because of her unwillingness to accept suggestions. She also talked about having temper tantrums and throwing things around in her room. These tantrums were referred to in both interviews, a year apart. She has only a few friends. One of them, a Jewish girl, whom she admired very much, she was forced to desert on the insistence of her other friends.

Her vocational plans are undecided. In the tenth grade she expressed interest in history and archeology, and the next year in languages. She wants to go to Stanford University, however, because "the climate suits her health and the architecture her temperament." This is contrary to the wishes of her family, who want her to enter Bryn Mawr. She has had no vocational experiences.

Summer vacation activities include a trip to Mexico (subsequent interest in Spanish), summer high school work in Spanish, and the study of Italian by herself.

Recreational and club activities are limited in number and are mostly solitary in nature. Orchestra is the only club activity in school, which is less than average for high school students. Athletic experiences include riding, swimming, cycling, and walking. She hates and fears "gym." She listens to the radio, reads, and attends a few movies, and confesses she does not know how to play. She reports that her health is good.

This record reveals several adjustment problems and their probable sources. There is a tendency to withdrawal and a certain degree of difficulty in adjusting to other people, both adults and those of her own age. These difficulties apparently have not been noticed by the classroom teachers. Her choices of free activities, which do not include many usually chosen by girls of her age, concentrate exclusively on solitary activities. She has few friends, and her relations with them are somewhat complicated. Immaturity is shown in her vocational plans and experiences. Her reasons for choosing a college seem far-fetched and affected. Part of the sources of her difficulties lie in her home life. At least the fact that she lives out of town, in a household composed of elderly adults, may be sufficient cause for her lack of contact with people her own age, and hence a cause for her apparent adjustment difficulties.

Summary of the P.E.A. Test Data

INTEREST INDEX, TEST 8.2a²

Category	Jane's		Class Median	
	Likes	Dislikes	Likes	Dislikes
Social Studies.....	38	0	51	13
Biology.....	19	0	56	13
Physical Science.....	25	0	56	13
English.....	75	0	63	13
Foreign Languages.....	100	0	63	6
Mathematics.....	0	25	43	25
Business.....	0	0	56	6
Home Economics.....	13	6	44	19
Industrial Arts.....	31	0	44	18
Fine Arts.....	88	0	38	12
Music.....	76	6	56	12
Sports.....	12	38	56	18
Manipulative.....	37	3	44	21
Reading.....	54	0	58	14
Total.....	39	6	52	21

In the twelfth grade as well as during two previous years, Jane's interest pattern is highly selective. Strong preferences are shown in four areas: English, fine arts, foreign languages, and music—foreign languages being the highest. These choices reveal two types of basic preferences: verbal activities and creative activities. The areas having to do with life realities, practical activities, and precise thinking are conspicuously lacking in her pattern of *likes*. The general tone of her interests in areas other than the ones mentioned above is that of indifference. Thus, the activities classified as biology, physical sciences, home arts, business and sports are a matter of indifference to her. Her total "dislikes" comprise only 6 per cent of all of the items.

² For a detailed description of this test and of the meaning of the summary categories see p. 338, Chap. V.

In the area of sports, however, she shows marked negative responses. Her *dislikes* here are in the highest quarter in the class. This is significant, because Jane has few *dislike* responses. Her remark to the counselor about her fear of gym corroborates this evidence but offers no explanation. Considering the fact that her choice of free recreational sports activities is limited to solitary activities, and also the fact that there is no evidence of a physical handicap or lack of physical skill, one is inclined to suggest that her negative reaction to sports occurs at the points of group or team activities. There is also other evidence suggesting that she dislikes and avoids activities involving social or competitive contact. Thus, on a previous questionnaire she showed very high dislikes on items concerning leadership and sociable activities. One is also reminded of her remarks to the counselor to the effect that she could not work or play with other people.

From these facts one develops an hypothesis of a solitary girl with a rather concentrated and somewhat narrow range of interests, which deviate in many aspects from the average pattern for girls of her age. An interesting inconsistency is apparent in one spot. Her score on interest in art is high. Yet her activity record shows no participation in art activities. Her lack of participation in art activities in the school might be due to the fact that her school schedule did not permit it, but she chose a second foreign language rather than art as an elective, and a study period rather than an art club. Neither is there any hint of artistic expression among her summer activities. On another questionnaire she shows no special preferences except in instrumental music. As will be seen later, her responses to free reading do not include a tendency to translate impressions gained from reading into art expression. It may be that her "art" interest is entirely passive, or that this interest is "spurious" in the

sense of being a symbolic expression of some other difficulty or problem.

Free Reading and Cultural Activities

Another series of data on her interests and preferences comes from her free choice activities and reading record. She reads the local daily paper and occasionally the *New York Times*. This latter, she says, is her favorite paper because of the book, art, and music notes. She is, however, unaware of the political theory favored by the papers she reads. (This is rather common, though, among high school students.) She spends an average amount of time (four hours per week) reading newspapers. Interesting, however, are the items she remembers from her reading during one month. These deal mostly with music (death of Chaliapin, opening of Robin Hood Dell) and international news (quake in Mexico, Hungarian countess married, taking over of American oil interests by Mexican government, Señora Cardenas and her friends giving their jewels to help United States oil interests, former Ethiopian ruler paying back dues to League of Nations). There are no items of national importance among the list of items she remembers, nor does she pay any attention to the editorials.

Her free reading during one sample period of a month (May 6 to June 6) included the following books: Wilder, *Bridge of San Luis Rey*; Wallace, *Fair God*; Lewis, *Charles of Europe*; Sabbatini, *Stalking Horse*; Ellis, *The Soul of Spain*. These are books about countries other than the United States, or by foreign authors. Her reading over a period of a year is twice as voluminous as the average for the class. Her magazine reading is rather average in quantity and character. Thus the *Ladies' Home Journal*, *Saturday Evening Post*, *Time*, and *Woman's Home Companion* are read regularly, mainly because they are received at home. The only deviation from the usual pattern is the reading of the

National Geographic regularly and in full, and the omission of the *Readers' Digest*. *National Geographic* was subscribed for at her request. At no time has she made use of the periodicals in the school library.

She attends no concerts, which is surprising in view of her apparent interest in music, of her proximity to a major orchestra, and the tradition in the region of attending concerts. She has attended no plays. She spends a lot of time, though, listening to music over the radio, her favorite programs being Charlie McCarthy, Ford Sunday Evening Hour, RCA Magic Key, Radio City Music Hall, and La Rosa. Archery is her only other recreational activity.

All of this is rather consistent with what was suggested by previously given facts about her interests and personality pattern. The impression of her preoccupation with the far-away and the esoteric is reinforced by her reading selections. Her failure to face the "here and now" is again emphasized.

APPRECIATION OF LITERATURE, TEST 3.3³

Category	Jane's Scores	Class Median
Likes Reading.....	100	62
Wants More.....	60	75
Curious.....	100	55
Expresses Other Media.....	35	25
Identifies Self.....	50	60
Relates To Life.....	100	70
Evaluates Reading.....	100	70
Totals		
Appreciation.....	84	65
Non-appreciation.....	15	40
Undecided.....	1	1

With the information about her reading interests at hand,

³ For a detailed description of this test and the meaning of the summary categories, see p. 253, Chap. IV.

it is interesting to look into her responses to free reading. The results from this test conflict at several points with the impressions from data up to this point. She apparently likes reading very much and is also curious about the background of authors and of the settings of literary works. This is consistent enough with her voluminous reading. She shows a much higher than usual tendency to relate what she reads to life and to evaluate reading, which is surprising in view of her apparent lack of interest in matters concerning life realities. As will be seen later, however, she shows little ability to discriminate between what is true to life and what is not. It has already been noted that while she has indicated high interest in the arts, she does not show any strong inclination to translate her impressions from reading into other art forms. People who are withdrawn and rely much on reading to secure experience with life are usually inclined to respond to reading with a high degree of self-identification. This is not the case with Jane. Her score on identifying herself with what she reads is below the median of the group and also below the usual scores in the same grade. This may, however, be a mark of sophistication in reading.

CRITICAL-MINDEDNESS IN THE READING OF FICTION, TEST 3.7⁴

	Judicious	Hypercritical	Uncritical	Uncertain
	%	%	%	%
Jane's Score.....	40	36	33	25
Class Median.....	70	18	22	5

According to the results from this test, Jane is not very successful in distinguishing realistic life situations from the dramatic or melodramatic ones. Her recognition of lifelike situations (judicious decisions) is the lowest in her group.

⁴ For the description of this test, see p. 266, Chap. IV.

She also has a strong tendency to be hypercritical: to judge situations and behaviors which are usually considered true-to-life as the opposite. These data support the impression of her lack of experience with life realities, and her immaturity in dealing with them. At many points she finds it impossible to make up her mind. This test is not good enough to be conclusive on this point, but it gives rise to some doubt about her literary judgment, in spite of her voluminous reading and her high score on disposition to evaluate reading.

INTERPRETATION OF DATA, TEST 2.51⁵

Category	Jane's Scores	Class Median
General Accuracy.....	54	57
Accuracy with Probably True and Probably False.....	35	38
Accuracy with Insufficient Data....	51	58
Accuracy with True and False.....	76	73
Overcaution.....	48	21
Going Beyond Data.....	43	36
Crude Errors.....	11	8

In techniques of getting meaning from quantitative data requiring precise thinking, Jane is near the average for her class. Her scores on accuracy are slightly below the median. This indicates inability to recognize the limitations of data. An examination of types of errors shows a greater than average tendency to go beyond the data, or accept generalities ignoring the limitations of the data. Not only is this score among the highest in the class (significant, since most of her scores are close to the median), but the proportion of errors in this direction in comparison to those in the direction of overcaution is also larger than that of the class (Beyond Data: Overcaution = 43:18, Class = 36:21). Her score on crude errors is one of the highest in the class.

⁵ For a detailed description of these summary categories, see p. 51, Chap. II.

In view of her fairly high accuracy in determining the absolute truth or falseness of inferences, her inaccuracies in judging trends and probabilities may have been the result of somewhat careless reading, particularly in view of the previous hints of difficulty with details requiring precise work and application, such as low scores on mathematics and difficulties in areas where detailed application and precision was demanded. However, there is strong enough evidence that Jane does not have the techniques necessary for precise manipulation and judgment of trends. There is also sufficient evidence that in instances where she does not get accurate meaning from the data, her tendency is to overgeneralize rather than to undergeneralize. The possibility of lack of

ABILITY TO APPLY PRINCIPLES OF LOGIC, TEST 5.¹⁶

	Jane's Scores	Class Median
Definitions		
Right Conclusions.....	6	4
Right Reasons.....	2	2
Total.....	8	7
Indirect Argument		
Right Conclusions.....	0	4
Right Reasons.....	0	0
Total.....	0	4
Ridicule		
Right Conclusions.....	6	6
Right Reasons.....	5	3
Total.....	11	9
If-Then		
Right Conclusions.....	4	2
Right Reasons.....	0	0
Total.....	4	2
Total		
Right Conclusions.....	16	18
Right Reasons.....	7	5
Total.....	23	22

¹⁶ For the description of this test, see p. 115, Chap. II.

experience is ruled out on the grounds that while the class improved over the period of one year, Jane's pattern of scores showed practically no change. Apparently the experiences provided for the class did not "take" with Jane.

Apparently Jane's ability to apply principles of logic, such as the importance of definitions in arriving at conclusions, the recognition of the limitations of indirect argument, the fallacy of trying to disprove by attacking the opponent, and the logical necessity of accepting conclusions flowing from the assumptions one has accepted, is approximately at the average for the class. Her highest score is on recognizing the futility of ridiculing the opponent as a method of argument. Her lowest score is in recognizing the limitations of indirect argument in proof. She seems to use "common sense" logic but is not particularly conscious of the principles she applies and has not developed finer techniques of reasoning. Since the class had devoted a good deal of attention to applying principles of logic of this sort, the cause must be sought not in lack of experience but in lack of interest or ability. Apparently the ability to abstract from the concrete situation which is required in this test and to draw refined logical distinctions is not the strong point in Jane's intellectual make-up.

Jane's ability to recognize the logical relationships in arguments and to discriminate between relevant facts and assumptions and irrelevant ones is at the average for her group. However, since in each of the categories—relevance, support, criticalness—the number of reasons she attempts is considerably higher than the number of reasons she gets right, a tendency to a broad and somewhat indiscriminate reasoning is suggested. (The same tendency was manifested in her methods of interpreting data.) Thus while the actual score of "rights" in each case is at the median, she uses a large number of irrelevant considerations, avoiding the outright inconsistent and non-critical considerations. Thus, general common sense combined with the lack of precise techniques and cautiousness is again indicated.

NATURE OF PROOF, TEST 5.21⁷

	Jane's Scores	Class Median
General Accuracy.....	129	128
Relevancy		
No. Marked.....	96	76
Relevant.....	70	69
Irrelevant.....	16	6
Support		
No. Marked.....	66	48
Support.....	42	42
Contradict.....	8	2
Irrelevant.....	16	3
Criticalness		
No. Marked.....	30	22
Critical.....	20	19
Non-Critical.....	3	3
Irrelevant.....	7	2
Conclusions		
Accepts.....	5	6
Uncertain.....	3	2
Rejects.....	1	1
Qualifications		
No. Marked.....	16	17
Accurate.....	10	11

Apparently Jane's logical abilities are not very high. She seems to fall short on precise techniques in both inductive and deductive thinking. Her confession of depending on her quick grasp and on a last minute rush to complete her assignments suggests that throughout her career in school Jane may not have taken the opportunity to cultivate precise methods of thinking and handling facts. The concentration of her interests in the direction of the arts, requiring imagination, and languages, requiring memory, may have in addition militated against cultivating these processes.

⁷ For the description of this test, see p. 131, Chap. II.

APPLICATION OF PRINCIPLES IN SCIENCE, TEST 1.3⁸

	Jane's Scores	Class Median
General Accuracy.....	-9	18
Conclusions		
Attempted.....	12	13
Right.....	2	7
Reasons		
Attempted.....	12	18
Right.....	3	10
Unacceptable Reasons		
Technically False.....	7	1
Irrelevant.....	0	2
False Analogy.....	2	1
Common Misconception.....	2	2
Assuming Conclusion.....	1	1
False Authority.....	0	0
Ridicule.....	0	0

Jane is extremely weak in the knowledge and use of science principles. On this test requiring application of scientific principles to everyday problems, Jane's general accuracy is the lowest in the group. Although she attempted a total of 12 conclusions, only two were right while ten were wrong. Both of these scores are the poorest in the group. Similar behavior is shown in her use of reasons. Since the score on false principles is the highest among her unacceptable reasons, her chief weakness is ignorance of these principles, but this does not explain her failure to recognize her own limitations, and to avoid marking reasons which she did not understand. Lack of experience in science would ordinarily explain part of the difficulty, but the school record shows that Jane took general science in the tenth grade, which is usually sufficient to permit a better record on this test. One could therefore conclude that it is Jane's own aversion to or inability in this area of thinking that is at the bottom of

⁸ For the description of this test, see p. 84 ff., Chap. II.

her weakness. The school record shows that Jane was scheduled for a special course in general science in her senior year to give her more experience in techniques of precise think-

SCALE OF BELIEFS, TEST 4.21⁹

	Jane's Scores	Class Median
<i>% Liberalism</i>		
Democracy.....	73	69
Economic Relationships.....	84	38
Labor and Unemployment.....	76	74
Race.....	94	70
Nationalism.....	96	70
Militarism.....	87	70
<i>% Conservatism</i>		
Democracy.....	12	17
Economic Relationships.....	0	20
Labor and Unemployment.....	18	10
Race.....	0	6
Nationalism.....	4	12
Militarism.....	3	12
<i>% Uncertainty</i>		
Democracy.....	15	12
Economic Relationships.....	16	28
Labor and Unemployment.....	6	12
Race.....	6	10
Nationalism.....	0	15
Militarism.....	10	13
<i>% Consistency</i>		
Democracy.....	65	75
Economic Relationships.....	85	80
Labor and Unemployment.....	76	88
Race.....	90	80
Nationalism.....	90	77
Militarism.....	76	80
<i>Totals</i>		
Liberalism.....	83	65
Conservatism.....	7	15
Uncertainty.....	10	16
Consistency.....	77	77

⁹ For the description of this test, see p. 215, Chap. III.

ing—a special concession and departure from general policy. However, there is no report of Jane's achievement in that course nor are science information tests included among the standardized tests given. Thus, the reasons for Jane's difficulties with scientific reasoning remain obscure.

A glance at the picture of Jane's performance on various aspects of thinking in comparison with her achievement on information tests opens up an interesting hypothesis. As a student of high verbal ability and good memory, has Jane been permitted to exploit these two qualities without a sufficient challenge to other intellectual processes?

Two tests give data on Jane's social attitudes. One of these attempts to diagnose generalized social beliefs. Jane apparently has a clearly thought out pattern of social beliefs. Her scores on liberalism are high and evenly distributed over all of the six areas included in the test. Thus, she tends to approve government control on behalf of the general welfare, and to reject economic individualism. She accepts equality for Negroes and thinks they have the same qualities as white people. She favors the international viewpoint, a logical counterpart of her interest in foreign cultures. There are very few items to which she has responded in a conservative direction. She also seems to be rather certain about her beliefs. Her responses are highly consistent in all areas, though in one of them, democracy, she falls in the lowest quarter for the class, because the class has an unusually high level of consistency.

There is also a marked growth in her social beliefs from the previous year. At that time she was highly uncertain and inconsistent in all areas except in the area of nationalism. Social attitudes seem to be the only area in which Jane has made a greater growth than the group of which she is a member. One would judge, then, that Jane's social beliefs are mature and clear and probably arrived at by her own efforts.

SOCIAL PROBLEMS, TEST 1.41¹⁰

	Jane's Scores	Class Median
Comprehensiveness		
Total Courses of Action.....	6	6
Total Reasons.....	48	46
Accuracy in Reasons.....	31	33
Ratio.....	5.2	5.1
Confusion of Implications		
Number Inconsistent.....	1	4
% Inconsistent.....	3	9
Undesirable Reasons		
Untenable.....	9	9
Irrelevant.....	7	7
Dominant Values in Courses of Action		
Democratic.....	4	5
Undemocratic.....	0	0
Compromise.....	2	2
Dominant Values in Reasons		
Undemocratic.....	3	5
Democratic.....	24	26

In test 1.41 the task is to apply social values to controversial social problems. Here, also, Jane shows a preponderantly democratic outlook. Sixty-two per cent of the tenable reasons she has used to support the courses of action she chose are what are defined as democratic values. She applies them consistently, only 3 per cent of her responses being contradictory to the courses of action she chose. She also shows a higher degree of cautiousness here than on any other test. Thus, a larger than average fraction of the reasons she attempts are applicable to the courses of action she chose. The range of the implications that Jane sees is average for the group.

Apparently Jane does much better with forms of reasoning involving broad generalizations and general logical dis-

¹⁰ For a description of this test, see p. 180, Chap. III.

tinctions. One is also impressed and surprised by the coherence of her social outlook in comparison to the apparent immaturity of her personal philosophy and her personal goals.

SKILL IN USING LIBRARIES AND BOOKS, TEST 7.2

	Jane's Score	Class Median
References		
Right	12	
Wrong	9	
Score	15	24
Library Classification		
Right	6	
Wrong	4	
Score	8	14
Card Catalog		
Right	9	
Wrong	1	
Score	17	17
Reader's Guide		
Right	1	
Wrong	2	
Score	0	9
Index Information		
Right	8	
Wrong	0	
Score	16	13
Parts of Book		
Right	7	
Wrong	3	
Score	11	16
Information		
Right	6	
Wrong	4	
Score	8	9
Total Score	75	99

In skills in the use of libraries and books Jane shows marked weaknesses. Except for her knowledge of the card catalog and the use of index information, in which she is at

the median for the group, she shows marked deficiencies, particularly in knowledge of the use of the Reader's Guide. Her total score is the lowest in the class. Again a deficiency with techniques of work is indicated.

By way of general summary, one may point out that Jane has good general ability, particularly verbal ability. She has a measure of success in logical thinking, but falls down in all areas requiring precise knowledge, precise processes of thinking or precise skills. In some respects her techniques of work seem quite deficient. Her social attitudes are mature and liberal. Her interests are highly selective and concentrated on esthetic pursuits, with preference for passive rather than productive activities.

Deficiencies and difficulties seem to be greatest in the area of adjustment to other people, both adults and age-mates. She seems immature in her attitudes toward herself, other people, and work. Her personal goals and ambitions are fanciful and show little thoughtfulness.

Apparently she has had altogether too meager an experience in challenging, concentrated work, and has cultivated a tendency to take her work and to approach her interests somewhat lightly.

It is difficult to tell what would have happened had the faculty become cognizant of her difficulties sooner. The faculty made several efforts to meet her needs during her last year at school. Arrangements were made to send her to college away from home (neither Stanford nor Bryn Mawr) with the proviso that she live in the dormitory. Special science work was arranged in an effort to give her training in precise thinking. To prevent her being lost in a large crowd, she was shifted from a large orchestra to a small string ensemble, and from mass hockey into a smaller archery group, in which she "made the team." Her further progress can only be traced in reports on her work in college.

METHODS OF INTERPRETING AND USING EVALUATION DATA *For Guidance of Individual Students*

As was described in Chapter I, one important purpose of evaluation is the guidance of individual students. The techniques of interpretation illustrated by the case study were especially relevant to this purpose. First, the meaning of the separate scores had to be clearly understood. The names given to these scores, such as "comprehensiveness," might be misleading unless related to the behavior required by the test. The meaning of these scores was further determined by their deviation from the group average as well as the level of expectancy for a given student.

Second, scores on any test were examined in relation to one another to arrive at a central pattern of behavior. In several instruments the scores were so dependent upon one another that the meaning of any one of them was not clear until the others are examined. For example, in the Scale of Beliefs two students might both have a score of 50 on liberalism, and one might say at first that they were equally liberal. But if the first had a score of 40 on conservatism and 10 on uncertainty, while the second had a score of 10 on conservatism and 40 on uncertainty, it is apparent that they were not equally liberal. The first had made up his mind on 90 per cent of the issues presented in the test and divided his opinions almost equally between the liberal and conservative viewpoints. The second had made up his mind about only 60 per cent of the issues, but his liberal responses predominated in the ratio of five to one. He was thus far more liberal than the first student, although his score on liberalism was the same.

There were even occasions when the interpreter had to be aware of the possibility of a considerable shift in the original meaning of the score, when that score appeared in certain combinations with other scores. Thus a high score on crude errors in interpreting data (marking true statements as false

and false as true) usually indicated a lack of even rudimentary skill in drawing inferences from data. If, however, the scores on accuracy were high and scores on other types of errors low, this score indicated careless reading of qualifying phrases in the test statements, rather than a deficiency in techniques of interpreting data.

Interpreting a comprehensive set of data from a battery of tests and other instruments presented a still more complex task of relating variables and revising the meaning of each aspect of behavior in terms of the larger pattern. Thus, since interests and social attitudes were known to influence thinking, data on thinking needed to be examined in the light of evidence on interests and attitudes. Formulation of tentative hypotheses of explanation usually helped sharpen the examination of evidence that might be thus related. In formulating these hypotheses the interpreter was first assisted by the structure of the instruments presented in this report, for they were designed to reveal relationships between different types of behavior as well as possible causes of deviant behavior. Thus the tests of clear thinking provided some neutral, scientific problems and other problems in areas involving personal values and beliefs. If errors in reasoning were concentrated in the latter, the tests of attitudes and interests might show that the difficulty lay in lack of interest or in prejudice rather than in techniques of thinking.

Familiarity with common patterns of behavior in the school threw further light on the behavior of individual students. An ambivalent pattern of social beliefs might be only the result of conflict between the values emphasized by the school and those held by the community, and therefore might not be very significant in individual cases. If, however, the usual pattern of social beliefs in the school lay in one direction while an individual's pattern lay in another, this is significant for individual guidance. Similarly, if dislike of writing was prevalent throughout the school due to overemphasis on

written assignments in all classes, even a moderate exception to this general rule assumed significance.

This sort of interpretation was essentially a process of postulating several alternative hypotheses to explain deviant behavior, and of checking each hypothesis against other data to see which one was most likely to be correct. Once the most probable causes of important weaknesses were located, it was a problem for the counselor and the school staff to decide how serious the difficulty was for a given individual and what, if anything, needed to be done about it. Illustrative guidance procedures have been suggested in connection with each instrument as well as in the case study. Individual variations were too great to permit a comprehensive account of all possible constructive methods. The results of a consistent program of evaluation over a period of years suggested that certain methods work better than others in similar cases. However, it must be remembered that evaluation data alone could not solve the problems of teaching or guidance. They only provided a more adequate basis for solving them. Teachers were sometimes annoyed when a program of evaluation revealed certain weaknesses in their program or in some of their students without indicating precisely what was to be done about those weaknesses. They sometimes concluded that the tests were useless. This is like saying that a thermometer is no good because it does not tell what to do about the weather. Tests could not be expected to solve all the problems of education, but they could and did call attention to many of the problems to be solved.

*For Checking the Effectiveness of
Curriculum in Achieving Major Objectives*

Another important purpose of evaluation was to discover whether the school was achieving its major objectives. Most schools wanted to develop citizens who could think clearly, who had democratic social attitudes, who were well adjusted, and the like. Evaluation data indicated the degree to which

changes of this sort were taking place. For this purpose interpretation of group data was necessary.

In the main, the processes employed in interpreting group data were similar to those employed in examining data on individuals. In each case it was necessary to determine the meaning of individual scores by reference to a more general pattern. In both cases hypotheses formulated at any point were checked against further evidence.

The usual method employed in locating strengths and weaknesses of a whole group—namely, considering the averages and the distributions of scores—was used with these data. By this method it was possible to determine the status of the group in the separate aspects of behavior measured by each instrument, such as the ability to distinguish facts from assumptions, or the tendency to mistake popular misconceptions for sound scientific principles. Frequently, however, it was necessary to determine also which *combinations* of behavior were common to many students in a group, thus requiring a common treatment.

Thus in the case of interests, the recurrence of a combination of high interest in music and art, or a combination of negative responses to English, reading, and foreign languages by many students were important kinds of evidence for diagnosing the group. Group medians and distributions of scores in each of the separate categories did not yield evidence of this type. A comparison of the patterns of interests of all individuals in the same group was needed.

Three types of processes were usually involved in estimating the progress of a group: A comparison of the scores by groups in the same grade or by successive grades in the same school, a comparison of scores made by groups in other schools with a comparable curriculum, and a comparison of student achievement with the behaviors specified in the statements of the objectives.

While the only satisfactory measure of growth was the

record of the same class over a period of years, a rough indication of the success of a school program was secured at once by comparing scores on the same test for successive grades. In some areas of objectives, the median of each grade was considerably higher than the median of the preceding grade, while in other areas, there was no significant difference in the grade medians. While the latter might be the general picture, particular classes taught by one or two teachers made significant progress. It then became the duty of the school to discover the factors which could account for the difference.

The most convenient method of comparing these scores with scores made by comparable groups in other schools might have been with reference to national norms. Thus, while progress might be shown from grade to grade on the test of interpretation of data, the median of each grade might stand in the lowest quarter of scores made by all other pupils of this grade who took the test. Unless some special factor was at work, such as very low reading test scores for the school population, this might indicate at once that still further progress must be made before the school's record could be considered satisfactory.

This method, however, was avoided as much as possible in the Eight-Year Study for several reasons. In the first place, it was recognized that as long as there were important differences in objectives and curriculum practices among schools, it would be inappropriate to measure progress by the same standards, particularly if these standards represented nothing more than an average of the performance of different groups under varying circumstances. The pattern of interests in a school for foreign students in New York City could not necessarily be considered appropriate as a "norm" or desirable pattern of interests for a suburban school in the Middle West, and the average of the two patterns might not be desirable for either school. Similarly, one would not expect students in a school which was barely beginning to explore the

methods of developing critical thinking to be judged by the same criteria as were students who have had long and careful training.

Difficulties were also encountered because of the methods of using norms to which teachers had been accustomed. The national average had been invested with almost magical significance, so that many teachers were too easily satisfied if their groups came up to it, even when they might have greatly exceeded it, and too easily discouraged if their groups fell below it, even though their progress was all that could be expected. For this reason, only tables of medians of comparable groups in other schools were made available to the evaluation representatives of schools in the Eight-Year Study, who were trained to interpret them. These gave a rough and admittedly cumbersome method of estimating the relative progress of comparable groups, but it was hoped that by this very fact a more thoughtful use of norms would be stimulated.

A third possible method of interpreting scores to indicate the success of a program in reaching its objectives was a comparison of the level of ability revealed by the tests with the level of ability required in life situations. Thus, if the use of the correct scientific principles in life problems were the objective of the school, and the tests revealed that students accepted a variety of popular misconceptions as scientific principles, then the school had not done enough in this direction, even though all other schools showed a similar weakness.

This sort of interpretation, however, had always to be made cautiously, because the level of accomplishment demanded by life situations was often a matter of vague conjecture. It was thus easy to expect too much or too little of students. The present level of achieving these newer intangible objectives may be too much determined by inadequate methods of helping students achieve them. Nevertheless, some comparison of pupils' performance with life demands was in-

escapable if we were not always to rest content with what other schools were doing. Perhaps none of them was doing enough.

For Checking Hypotheses Underlying the Program

A third important purpose of evaluation was to check the hypotheses underlying the school program. Often new practices were introduced in the hope of producing certain desirable changes in students. These changes might not come about, or they might be accompanied by other changes which were less desirable. One public school introduced a core program with several purposes in mind, one of which was to develop better social attitudes. A comprehensive testing program revealed that while the social attitudes developed were clearer, more consistent, and more liberal than in most schools in the Study, the students had serious difficulties with techniques of precise thinking. In drawing inferences from data, they exhibited little caution and showed a tendency to go beyond the data. In applying facts and principles they failed to discriminate those which were valid and relevant from their opposites. Apparently in emphasizing social values the school relied too much on generalizations and too little upon the careful analysis of factual data.

In another school the evaluation of reading revealed that one group specializing in science and mathematics showed a more limited appreciation than all others, including those in other grades specializing in the same field. They found little enjoyment in reading; they did not identify themselves with their reading or relate their reading to life problems. Since this was a marked deviation from the type of responses prevailing in the school, the problem was considered by the faculty. It developed that a special course in literature was offered to this group. On the hypothesis that science students are interested in scientists, this course concentrated on biographies of scientists and mathematicians. Since it was not the

intention of the staff to narrow the reading interests of these students, a broader program was agreed upon.

Still another school had hoped to develop democratic attitudes by means of a program of extra-curricular activities organized by the student council, while conducting its academic curriculum in the usual manner. The results of the test on Beliefs about School Life revealed that a large majority of these students preferred authoritarian methods of classroom management, approved of social distinction of all sorts, and in general had tendencies toward undemocratic attitudes. These results called into question the efficacy of this program of student activities for the purpose of democratizing school life. In the course of an investigation by a group of students and faculty members, it was discovered that the student council was run by an inner clique. Many of the student activities tended to be exclusive and to have other undemocratic characteristics. The active participation was limited largely to students in the upper grades. In the light of the facts brought out by this study, a reorganization of student activities was undertaken, involving a closer relationship between curricular and extracurricular activities.

Such instances indicated that special care had to be exercised when changes were introduced into the program to find out not only whether the intended results were produced but also whether undesirable features did not accompany them. Even if no major changes had been made, the hypotheses on which the school had always operated might be faulty. Hence, evaluation data needed to be examined with special reference to the *issues* underlying the program.

Possibility of Interpretation

The foregoing discussion may have left the impression that interpretation of evaluation data required very unusual insight and patience, and too extensive knowledge of evaluation for the classroom teacher to master. There is no getting

around the fact that a thoughtful interpretation of the evidence on students' progress and the effectiveness of curriculum practices is complex, and that it can be learned only by long practice supplemented by careful explanation. Yet there is no reason to believe that further progress in getting a more adequate picture of pupil growth will ever return to the primitive simplicity of school marks. Reducing the amount of data secured is no solution, for a few scattered data can only raise questions, not answer them. A rich and full program of evaluation can suggest answers to a great many questions, but only by thoughtful interpretation and not by chance. Teachers must learn to get meaning from the extensive and well-integrated sets of data now available. Unless somebody knows what the scores mean and takes them into account in his teaching, it is obvious that there is no point in getting them.

On the other hand, the process of interpretation is not so difficult for busy teachers in a large public school as the foregoing may suggest. When teachers know the pupils concerned, hypotheses to account for their test scores readily occur to them. Then, too, if evaluation is carried on continuously, the evidence accumulates gradually, and only a few data need be interpreted at any one time, and fitted into what one already knows about students. Also, the processes which appear elaborate, when written down and explained verbally, easily become part and parcel of the common sense thinking of thoughtful teachers. Finally, when evaluation is undertaken as a common task for the school, with the whole faculty cooperating in interpreting the results, the task for any one individual is reduced.

Chapter VIII

PLANNING AND ADMINISTERING THE EVALUATION PROGRAM

~~~~~

The preceding chapters have already dealt with many of the basic problems in planning and administering an evaluation program. They have discussed the purposes of evaluation, its basic assumptions, and the steps which must be followed in developing appraisal instruments. They have indicated an appropriate division of labor among teachers, school officers, and experts in evaluation. They have suggested a possible classification of school objectives by types of behavior, each of which requires a different technique of appraisal. They have described instruments and techniques for the study of growth toward objectives usually regarded as "intangible," such as certain aspects of thinking, social sensitivity, appreciations, interests, and personal and social adjustment. They have reported in great detail the method of construction of these instruments so that teachers might develop others. They have indicated, at least by implication, certain characteristics which are desirable in evaluation instruments developed or selected by a school staff. In addition to those usually discussed, such as validity, reliability, objectivity, appropriateness to age levels, and the like, the characteristics given special emphasis in this report were the diagnostic value of the multiple scores yielded by these instruments, and the interrelationships of these instruments, so that each score was supported and explained by other scores on the same or other instruments. Finally, the previous chapter dealt with methods of interpreting and using evaluation data.

All of these considerations are pertinent to the problem of

planning and administering an evaluation program. In addition, certain administrative procedures are essential to assure the comprehensiveness of the appraisal, to secure the cooperation of the entire staff of the school, and to increase the practicability of the program. When testing is left to each individual teacher, there is likely to be incoordination, and the most important objectives—those to which the whole school program is dedicated—are frequently overlooked, especially since they are usually the hardest to evaluate. Objectives which are easiest to evaluate may be emphasized out of all proportion to their importance and, as a result, attention may be drawn away from other equally important objectives. No data may be secured relevant to the hypotheses on which the school is operating. Pupils may be overburdened with tests in certain departments or at certain times.

If, on the other hand, the actual conduct of the appraisal is left to an evaluation specialist, there is the danger that pertinent data will not reach the teachers who should act upon them. The results may be reported in a form which teachers cannot readily understand, and recorded in ways which involve undue clerical labor. A most common defect is that all available time and effort are spent in gathering data, with none left over to interpret or use them for individual guidance or curriculum improvement.

It is the intention of this chapter to discuss certain principles and procedures of planning and administering an evaluation program which have helped to make it effective in some of the schools participating in the Eight-Year Study. For the sake of brevity, no account will be given of the gradual development of these practices, and only occasional references will be made to the diversity of practice on these points now prevailing among the cooperating schools. The chapter will attempt to describe a few illustrative practices in planning the program, recording the data, and providing for their effective use.

*Planning the Scope and Emphasis of the Program*

Early in the Study it was found that a comprehensive evaluation program required careful, cooperative planning by the staff of the school. The data necessary for a well rounded picture of individual development, of the progress of the group, and of the effectiveness of the curriculum would not be secured if the task was left to individuals. It was quite evident that the staff as a whole must decide what to evaluate, what kinds of evidence to secure, and how to go about securing evidence and using it. As the first step in evaluation involves the formulation of the school's objectives, this cooperative planning of evaluation began with this step. In order to secure a statement of objectives which was representative of the work done in the school and thus to make sure that no phase of growth really emphasized in the school was neglected, the whole staff participated in the process of formulating the basic platform of objectives. Each teacher or departmental group of teachers submitted a list of objectives. These lists were then considered by committees and by the whole faculty in order to clarify them further and to discover where there were common emphases and where unique types of development were indicated.

If there was any conflict between the appraisal of the school-wide objectives and those held by individual teachers, it was rather commonly assumed that the first responsibility of the school was to its general objectives. While the principle was never abandoned that the school as well as individual teachers should do all they could to study growth toward the objectives unique to the specific courses, the larger principle usually prevailed that the study of the most important aspects of human development as expressed in the general objectives should be the major concern of a school. The nature and extent of the appraisal of the specific objectives was to be planned so that it was consistently related to this general program and helped to support and clarify it.

Fortunately, the areas of objectives of general concern were usually limited in number and thus did not constitute too heavy a burden either on the resources of the school or the time and tolerance of the students. For example, most schools were concerned with one or more phases of critical thinking, social attitudes, certain work habits and study skills, interests and appreciations, social adjustments, and certain types of functional information. Hence, in most schools there was sufficient opportunity to carry on additional specific investigations of student growth.

A second major principle governing the planning was that appraisal was to be continuous. The adoption of this policy meant that the schools had to consider the time and effort needed for a continuous check before decisions were made regarding what range of objectives would be appraised, or how detailed the check was to be. As can be seen later, this consideration also determined the calendar adopted for the administration of the evaluation instruments.

It was also clearly understood that it was the program of the school and its effects on student growth and not the individual teacher or pupil that was being appraised. The effectiveness of evaluation is likely to be impaired if the evaluation program is conceived by the teachers either as an extension of the usual examinations and marks in courses or as a means of judging their competence. With the first misconception, teachers may try to find the strengths and weaknesses of their pupils with the idea of rewarding the strengths and penalizing the weaknesses, accompanied by some exhortation to do better, but without making any significant change in their courses, or still less in the whole school program. With the second misconception, teachers may try to justify the present situation rather than to seek fully and frankly for points needing improvement. For these reasons the schools favored instruments and devices which yielded descriptive diagnoses of students and which, because of this character-

istic, could not be easily converted into grades and marks. Most of the evaluation instruments used also diagnosed the kinds of behavior capable of development only through concerted and cooperative efforts of many teachers over a period of time, and not by the work of one teacher in one course or unit of work.

Finally, it was understood that the evaluation program was to serve the local needs and purposes of each school. The particular emphasis as well as the extent of the program was largely determined by what each school needed data for. Thus many schools had set up an experimental program on some central hypothesis. Checking that particular hypothesis often required a detailed appraisal of certain specified types of growth, such as in critical thinking, in range and maturity of interests, in social sensitivity. In these cases the evaluation program was planned to give most detailed evidence on these points. Local conditions also influenced the plans. For example, some schools drawing students from widely scattered places had to concentrate the evaluation in the earlier grades on the diagnosis of interests, abilities, and basic skills. Still other schools had differentiated sequences of programs, calling for evidence necessary for the placement of the students in these sequences as well as for determining the relative effectiveness of these programs. Often special effort was needed to appraise the acquisition of common skills in order to answer the questions of parents and the community who feared that the new curriculum might neglect these outcomes.

Certain practical considerations also limited the plans. While most schools made an effort to plan the scope and the nature of their evaluation programs according to what they thought to be important objectives or crucial needs of their programs rather than in terms of economy, immediate availability of instruments and techniques, or the ease of their administration, it was natural that the cloth had to be cut ac-



cording to the resources of the school. Thus, financial expenses were involved in administering the testing program even though much of the scoring was done at the evaluation headquarters. Someone's time and effort was required for handling the data, since there was no point in collecting more data than could be properly recorded, interpreted, and used.

In determining how to adjust the scope of the program to the limitations of resources, the general principle followed was to plan to appraise at least in limited fashion each of the major areas of objectives before planning a more detailed evaluation of a single area. This seemed wise first because it was recognized that evidence covering a fairly broad range of behavior is needed for proper appraisal of the program of a school. The schools also realized that teachers tend to emphasize the areas of development the results of which they can see more clearly. An even distribution of efforts of appraisal over the significant objectives was thus expected to produce a more even distribution of emphasis in teaching. Finally, since detailed appraisal was usually given to areas of objectives which were easiest to appraise or in which instruments were readily available, it seemed wise to make sure that some of the important intangible objectives for which no refined techniques or instruments were as yet available would not be overlooked.

Generally speaking, then, while the schools attempted to evaluate as broad a range of objectives as possible, the actual program rested on decisions representing a combination of the ideal possibilities and the practical limitations of the school situation.

### *Collecting Data*

Once the staff agreed on the general scope of the program, it considered the methods for securing the needed evidence. This required a preliminary survey of the data already available in the school. Only when the faculty had explored the

possible relationships to school objectives of the data which was already collected was it in a position to decide what further data were needed. In the process of clarifying the school objectives it was usually discovered that the faculty was already collecting many types of data on these aspects of development. Thus, many schools had a testing program including aptitude tests, reading tests, and information tests. Most schools also had an abundance of less formal types of data collected in the normal process of teaching and administering the school. In most cases these data were put to only a limited use, partly because they were scattered, partly because of the tendency to consider only the scores on objective tests as appropriate evidence, but mainly because their bearing on the objectives of the school was not evident.

When, however, the objectives were clarified to the point where teachers could clearly see the concrete behaviors involved, the bearing on the broader objectives of some data which teachers were collecting for specific purposes became apparent. Thus the English teachers found that student writing could be examined for evidences of interests, social adjustment, and social attitudes as well as of the ability to spell and write correctly. Records of free reading were found to yield evidence on maturity of tastes as well as of quantity of reading. Even such simple data as the records of activities and subjects taken assumed significance when considered in the context of other facts about the students.

This examination of the data already available usually indicated certain gaps, that is, objectives on which little evidence was being obtained. Hence, the next step was to plan the ways and means of securing the additional data needed. Usually at this point there was a tendency to consider only paper-and-pencil tests. However, a careful analysis of the methods of securing evidence most appropriate to each objective revealed that the classroom situations provided a far greater source for securing data on students than had usually

been assumed. For the appraisal of some objectives, such as the ability to plan the attack on research problems, or to use laboratory techniques and tools, the observation and recording of student behavior in normal classroom situations was the best if not the only adequate source. Thus, one school secured data on student growth in planning research by the simple device of providing students with pads on which to record in duplicate the successive outlines of the plans they made. At other points semi-controlled classroom situations, suitable for both learning and evaluation purposes, could well be used in place of formal tests. Thus the difficulties encountered in getting information from libraries and books could be diagnosed, and in many schools were diagnosed, by giving students assignments requiring the use of the library and by observing the methods they used in obtaining the necessary information.

These uses of sources of data in processes integral to teaching were found to be particularly helpful because when teachers were directly responsible for collecting evidence they more often used the results than when only the summary of data came to them. However, collaboration and systematic allocation of responsibilities on a school-wide basis are necessary to prevent this method from being too time consuming. In economizing effort it was found that certain departments or teachers of certain areas were in especially strategic positions to collect one kind of evidence, while others had greater opportunity to obtain information of a different sort. By systematizing the use of such informal devices and by making the results generally available, many schools found that they could extend the scope of their evaluation through the use of opportunities already existing in the classroom.

Having agreed upon the informal methods to be used in obtaining evidence, the next step was to plan the use of more formal devices. Usually paper and pencil tests were reserved for points where information was lacking altogether, or

where the available information was inadequate, or where the use of informal methods entailed too much time and effort. Thus, most schools had considerable evidence on information and skills, but little or no evidence on the growth of students in various phases of thinking. The information on social attitudes secured or securable through anecdotal records, classroom observation, or from student papers was found to be too scattered and meager to give an adequate picture of social beliefs over a range of social issues of importance. At many points, then, it was necessary to use additional paper and pencil tests, either because they represented the only appropriate method of getting the evidence or because they were most economical.

#### *Drawing Up a Schedule for Testing*

In setting a calendar for the testing program, it was necessary to consider several factors. In the first place, the total time devoted to testing could not be so great that students and faculty thought themselves overburdened with tests. To avoid this difficulty, careful estimates were made of the total time needed for taking all tests which were tentatively proposed for the program. Some schools even went so far as to set up a time limit and to eliminate certain instruments if the proposed schedule exceeded that limit.

In the second place, the schedule had to be drawn so that there was no undue concentration of formal tests toward the end of the year, and particularly toward the end of the twelfth grade, since such a congestion of schedule subjected students to unnecessary tension, and did not provide evidence at times when the results could most effectively be used. Generally, congestion was prevented by devising a tentative calendar for the testing program covering all the grades of the school. Such a calendar included the repeated use of certain instruments to check on growth as well as the giving of certain tests which needed to be used only once. Tests

yielding information basic to understanding new students and for the initial planning of teaching were usually placed early while others were distributed over successive years.

The schedule also provided for a fair distribution of time among the several subject fields so that the testing did not take an undue amount of time from any one class. This was done by allocating different tests to different departments in the school or by staggering the successive periods of the day used for giving tests.

The methods of organizing for this cooperative job varied greatly from school to school, depending on the size of the school and the make-up of their faculties. In some cases, particularly in smaller schools, the school psychologist or counselor took the major responsibility for drafting the tentative plans and for arranging the practical details. In such cases much of the participation of the faculty was achieved through informal contacts and personal conferences.

In other schools evaluation committees were established, whose responsibility it was to get the necessary information and advice from the rest of the faculty, to draw up a plan, and to care for the routines. Often members of such committees took special responsibility for giving certain instruments or series of instruments as well as for collecting certain materials from other teachers.

In still other schools the responsibilities were divided among the staff according to the types of evidence to be collected. Thus a psychologist became responsible for giving the psychological tests and reading tests. An evaluation representative supervised the use of the special tests developed by the Evaluation Staff, while individual teachers were responsible for information and skill tests in their respective areas. Whatever the particular scheme, it was found necessary to make careful, coordinated plans for the entire program of evaluation.



*Summarizing and Circulating the Results*

Since the evidence of student development was obtained from records already existing in the school, from collecting data easily obtained as part of the class work, and from especially selected tests and appraisal devices, the problem of organizing and summarizing these varied types of information was an important one. Part of the task of organization was accomplished by using a folder for each student, and placing all records relating to this student in this folder. The student folder became a file of information to which additions were made as the evidence accumulated.

However, the varied forms of evidence made it necessary to utilize additional techniques of organization. The test scores were already organized into patterns devised by the evaluation committees. In the case of data recorded by students or parents, such as entrance information, reading records, and written papers, the administrative problem was to organize the record keeping in such a way that a consistent and cumulative record became available. Thus, in case of the reading records, a certain time each week was allotted to students to write down the books they had read during the preceding week. Copies of written work were assembled in the student folder.

To obtain satisfactory records from observations made by teachers or other persons in a position to observe students involved several other administrative problems. Chief among these was that of obtaining observed facts on behavior, in place of ratings drawn largely from memory. Some organization was also needed to obtain a sufficiently representative sampling of the observations from different teachers, supposedly in a position to see the student in different situations. Staff conferences devoted to clarifying the behavior to be observed and the techniques of obtaining the record most economically, and then to periodic discussion of records sub-

to assume the task of summarizing this evidence and of passing these summaries along. Furthermore, they were expected to be most familiar with the tests relating to their objectives, hence they were usually expected to give these tests and to summarize the most pertinent points revealed in the test scores. If some other members of the staff, such as the psychologist, the counselor, or the evaluation representative, were responsible for parts of the testing, they assumed the responsibility for summarizing the results of the tests they gave.

These summaries of various items of data about a student were then brought together by the person mainly responsible for his guidance, usually his homeroom teacher or counselor. This person was responsible for making an over-all interpretation of the data, indicating the outstanding strengths and weaknesses, pointing out some factors contributing to these, and making some tentative suggestions regarding what needed to be done. Until this step was taken, one teacher knew about his language skills, another about his social attitudes, another about his techniques of thinking, another about his interests, but no one had a coherent picture of his development. Hence, few teachers were in a position to make comprehensive suggestions regarding what the student needed, or able to take constructive action.

While the summaries of specific data were usually made at the time when the evidence was secured and when the circumstances of securing it and its implications were fresh, the over-all interpretations were made only at certain regular intervals or at times when such information was most needed. That is, these interpretations were usually made at the times when reports to parents were being prepared and when particular curriculum plans were being made. From time to time the case of an individual student might require a special interpretation of his record. The members of the staff who made these over-all interpretations usually had some insight

into the psychological implications of behavior, some training in the interpretation of these types of data, and some personal contact with the students. In order that the data be actually used, it was found to be extremely important that all data on the growth of a student pass through the mind of a person who knew him and had a responsible relation to his all-round development.

The schools found it necessary to develop plans for circulating information as well as summarizing it. Usually the basic data collected by each teacher remained in his possession as long as he needed it. The summaries were, however, circulated as soon as they became available. This was done either by exchange of notes or by frequent meetings of small groups of teachers and advisers of each group of students. The latter method was most commonly adopted by schools where some form of core or unified curriculum was in force, in which case a small group of teachers was responsible for a major portion of the school experiences for a given group of students.

To facilitate still further the circulation of information, the basic files were placed in spots accessible to teachers and counselors. If there was a school counselors' office, the files were placed there. If teachers acted as counselors, their respective classrooms or offices contained those files. The main principle was to keep the records of students where they were most frequently used. Several copies were made of data which were needed in different places or by different people at the same time. Thus, often the basic entrance data were available in teachers' or counselors' folders as well as in the principal's office.

A somewhat different problem was involved in handling group data. It must be recalled that all data pertinent to individual growth could also be summarized so as to give evidence about the strengths and weaknesses common to groups of students. These group summaries were particularly useful

in appraising the effectiveness of the curriculum. Since the summarizing of group data requires a certain degree of statistical competence and since, furthermore, the analysis of these data involves comparative study of data on all groups in the school, these tasks were usually in the charge of a person or a committee responsible for coordinating the curriculum program in the school. It was the responsibility of this person or committee to analyze and summarize the data and to report periodically to the faculty on the effectiveness of the school program in achieving its major objectives.

The processes involved in interpreting group data have been described in the previous chapter. The chief administrative arrangement required was to provide time for the staff to meet together regularly to study the data, bringing to bear upon it the specialized competence and points of view of a representative sample of all departments in the school, and for cooperative planning of teaching. This time was usually secured by a more careful rearrangement of schedules and teacher responsibilities. A few schools reduced the total teaching period of the day by having students come half an hour later. In a great many cases teacher time was saved by teaching students to work independently and thus dispensing with teacher supervision at some points of their work.

#### *Using Evidence for Improving Teaching and Curriculum*

Availability of evidence alone, no matter how well organized and summarized, did not assure its effective use. The implications of the individual and group data to daily procedures in guidance, teaching, and curriculum making had to be intelligently digested by every teacher before the greatest value of the evidence was attained. It was necessary to make special provisions for teachers to develop the insight and techniques needed to translate into practice what was learned about the students.

Definitely scheduled opportunities to study the data was

one of these special provisions. To make maximum use of evaluation evidence in teaching and guiding students and in curriculum improvement was found to require continuous study and collective thinking by the whole staff. Occasional reports to the staff about the results of the evaluation program proved inadequate for this purpose. At best these occasional reports served only to acquaint the staff with the fact that something could be learned from the evaluation program. Similar limitations were found with occasional case study conferences regarding individual students. The occasional conferences introduced the staff to the techniques of analyzing evidence about individuals and suggested some possible implications, but they did not provide adequate opportunity for the staff to explore multiple explanations and to consider various constructive modifications in daily practices which were implied by the evidence.

A second provision was to see that the staff explored the evidence and its implications at those points where decisions were to be made and actions to be taken. When discussions of evaluation data took place apart from any need for action, they were often received by the staff with the passivity usually accorded to academic discussions and often regarded merely as an interesting theory. In many cases, what the staff seemed most to need was a clear demonstration of the helpfulness of the information to the teachers' ongoing activities. It was found to make an enormous difference in the attitude of the faculty toward evaluation data whether the data on a given student were just "studied" or whether they were introduced at a time when the staff was concerned with such questions as what to do about certain students' lack of success in academic work or apparent failure to adjust to the life of the school. Similarly, when such questions as the usefulness of Greek history for the non-academic students or the advisability of social mathematics for those failing in regular mathematics were raised, the evidence on the success



or failure of these groups in achieving various objectives assumed a greater significance. Not only were the implications of available evidence scrutinized more carefully, but the possibilities of constructive action were also considered more thoughtfully when the attack was made in terms of definite problems to be solved.

There were several occasions in the typical school procedure which proved to be particularly appropriate for demonstrating the usefulness of evaluation data and for initiating teachers into the habit of basing their decisions and practices on whatever evidence was available. Making out programs for the students for the year was one such occasion. Often, student programs were decided on the basis of such factors as: convenience of the time, college requirements, previous success or failure in various subjects, and the student's own wishes. When a fairly comprehensive set of evaluation data became available an attempt was made to reach these decisions in the light of all available data about the student. Frequently, also, the program making was done cooperatively by a faculty group in charge of a group of students. Such conferences served not only as a means of acquainting the teachers with what was in the "records," but also to clarify and unify the guidance policies of the school and as a means of initiating a habit of making decisions of all sorts in terms of evidence rather than in terms of previous practice or of unconsidered personal preferences.

Reports to parents offered another occasion to study the growth of students, to consider their needs, and to initiate the habit of making judgments in terms of evidence. Many teachers had felt at a loss in finding a sufficient number of valid things to say to each parent about the students. Examination of objective evidence proved to be very welcome at such times.

Most of the schools also had to consider from time to time certain changes in the curriculum. This afforded another

occasion for studying evaluation results. These suggested changes ranged from the proposal to add new courses to the possible reorganization of the whole structure of curriculum offerings. These occasions were an opportune time to survey the effectiveness of the curriculum in terms of available evidence. Several of the Thirty Schools began with occasional staff meetings considering such problems. They proved so useful that curriculum planning sessions held regularly through the year became a frequent practice. Many schools held prolonged sessions either in the spring after the school was out or in the fall before the year's program was begun. At this time the information about the growth of students toward all objectives of the school was carefully examined by the whole staff, and the curriculum plans as well as plans for teaching and special activities to be promoted were made in terms of that evidence. Weekly conferences throughout the school by smaller groups of teachers dealing with the same group of students were also a frequent practice.

A third provision was to involve the entire school staff in the study of the results of evaluation. Often consideration of the implications of the evaluation evidence suggested changes in practices which were not under the direct control of any one member nor any small portion of the staff. For example, in many cases the sources of difficulties in achieving consistent democratic attitudes appeared to be in the whole organization of the school, the weaknesses in clear thinking were apparently produced by an inconsistent approach among the different teachers, and adjustment problems could largely be traced to the way in which the program of student activities was organized. To uncover difficulties of this sort and to plan constructive remedies, it was necessary to take the whole staff into partnership in considering and formulating school policies and in examining evidence helpful in making wiser decisions.

As the evaluation program proceeded, it became increas-

ingly clear that to be effective it must involve extensive participation by the entire faculty. Teachers had to formulate objectives and to agree on the common objectives of the school. They had to select certain manifestations of growth toward these objectives which could be tested, observed, or recorded. While in the choice of instruments technical advice was needed, the final decision regarding their appropriateness rested with the teachers. Similarly, the final decisions regarding what was significant in the evaluation data and how they could be used in improving school practices could wisely be made only by those who were carrying on the job. When judgments and decisions of this sort were made by "experts" and passed on to the teachers, the results were less fruitful.

A program which involved wide participation naturally raised the question of the competence of the rank and file of teachers in such matters. Thus, for instance, the ability of teachers to interpret properly evaluation data, particularly those requiring psychological insight or technical manipulation, was questioned. The usual assumption, for example, was that only statistically trained people could be trusted to deal with test scores. The experience in the Thirty Schools was that on the whole teachers made better interpreters than persons statistically qualified but whose personal contact with students was limited. Moreover, since it seemed evident that unless teachers were trained to interpret evaluation data for themselves, their ability and insight in using the results as well as their willingness to do so would remain limited, the schools in cooperation with the Evaluation Staff embarked on the job of training the teachers for this work. Periodic conferences on interpretation were held in each school. To provide for more continued help, in each school one person was chosen to act as an evaluation representative and as an evaluation adviser to the rest of the school. This person spent some time receiving training in interpretation

either in workshops during the summer or with the Evaluation Staff during the school year.

Similarly, the use of evaluation data in shaping an improved school program could not be left to accidental or amateurish efforts. Some training and guidance of teachers was needed. This did not mean that all teachers were packed off to summer school to receive such training. Participating in planning and administering the evaluation program and in the study and application of its results in itself provided an opportunity for training hardly exceeded by any other device, provided there were opportunities in the school for the staff to think together on these matters and to make cooperatively decisions which had previously been made by individuals.

This brief report on the planning and administration of an evaluation program provides a further illustration of the ways in which the evaluation project was an integral part of the processes of teaching, of curriculum making, of guidance, and of teacher education in many of the Thirty Schools. As a result of its work with the schools, the Evaluation Staff is convinced that a program of evaluation can achieve its maximum usefulness only when it is an integral part of the major tasks of the school. Deriving its direction from the major objectives of the school, the evaluation program helps to clarify these objectives into clearly apprehended goals and purposes which are more effective guides to teaching and counseling. Exploring each major objective to identify types of behavior manifestations which will serve to reveal the progress of students toward this objective helps to focus attention upon the learner and the meaning of the educative process to him. Studying the results of evaluation serves to identify strengths and weaknesses of teaching and inadequacies in the school program. Effective participation in these several phases of evaluation serves as a stimulating experience for teachers in their own continuing education.





## PART II

### RECORDING FOR GUIDANCE AND TRANSFER



The work of the Committees on Records  
and Reports, and the Forms produced  
by them.

groups, for example, studied the objectives of teachers and schools, but always in relation to particular problems, and always with the results obtained by other groups available for comparison and use. The list of objectives prepared by the Evaluation Staff was particularly helpful to all the committees on recording and reporting.

All record forms that can do so provide space for information of the kinds obtained by the Evaluation Staff, so that this can be related to the other data and so can help to make a more complete description of the pupil.

Although it will be said again in relation to various forms and their use, it must be emphasized here that no single result of evaluation procedures or of observations recorded on the forms is considered to be independent of other information about a pupil. All the information obtained, as would be true if he were studied by a psychologist or psycho-analyst, contributes to the more complete understanding of him that becomes the basis for the school's dealing with him.

### *Philosophy and Objectives*

The original Committee on Reports and Records considered with great care former methods of recording facts about personal characteristics or traits, and the words used in describing and reporting about them.

Out of this study and the discussion of the problems facing the committee came the philosophy and objectives that governed the later work. The list of objectives in explicit or implicit form was reexamined by the other committees, and was generally accepted as a guide, though it was realized that some of it applied most completely to the study of personal characteristics.

### GENERAL PURPOSES AND PHILOSOPHY OF RECORDING

1. (a) The purpose of recording is not primarily that of bookkeeping. Instead the fundamental reason for records is

their value as a basis for more intelligent dealing with human beings.

The first purpose of records is therefore that of forming a basis for *understanding individuals* so that effective guidance can be given.

(b) Since the educational process is a continuous one that should not be set back at certain transfer points, it becomes necessary that guidance shall continue across such points in such a way as to increase the probability of continuity in dealing with the person.

An extended purpose of records hence becomes that of furnishing *transferable information* for guidance.

(c) Because of the need of cooperative and consistent dealing with a boy or girl by home and school, as well as the right of the home to information as complete and reliable as possible about progress and development, records should furnish the material on which reports can be founded, and reports should be considered an essential and consistent part of the recording system.

A third purpose of record keeping is therefore to provide the information needed for *reports to the home*, and to add effective ways of giving such information.

(d) Information is needed at all stages of education, and particularly at points of transfer from one institution to another, or from an institution to employment, in order that qualifications of the individual for the new experience can be fairly judged.

A fourth purpose of record keeping is therefore to provide information, and methods of transferring it to others, that will give evidence regarding a pupil's *readiness* for succeeding experiences. This would apply to fitness for a particular college or other institution.

2. What might be considered an indirect but nevertheless important purpose of records is that of stimulating teachers to consider and decide upon their objectives, judge some-

thing of the relative importance of their aims, and estimate their own work and the progress of their pupils in relation to the objectives chosen.

Many teachers think almost entirely in terms of the most obvious objectives concerned with the learning of subject matter and evaluate their results only in terms of such aims. They give little or no consideration to the changes in their pupils that should come about as a result of the experiences undergone, and so they fail to bring about the development that is possible. Through well planned records they can be helped to a wider vision and a more constructive influence.

It is evident that the most valuable and complete record that could be made by observation of an individual would consist of a record of his behavior throughout life, or that portion of it under observation. It is believed that any observational technique that has value must consist in using the parts of such a record that can be collected and arranged in the time at a teacher's disposal. This can be done by recording significant incidents of behavior and interpretations of them (the "anecdotal" method), by characterizing in one way or another the kinds of behavior observed (sometimes called "behavior description"), or by a combination of characterization and of supplementary analysis in paragraph form.

Where a teacher deals with a small number of pupils, or carries a light schedule, the recording of extensive anecdotal material seems possible and highly valuable. Some institutions and teachers use such a method even when the written material cannot be extensive. The more the demands on the teacher through appointments or pupil load, the less is it possible to write voluminously, and the more does it seem necessary for each instructor to digest his observations into quickly recorded (but not too quickly arrived at) judgments about the typical behavior of the pupils. No "checking" sys-

tem, however, can fit all of the significant differences among people, no matter how well it is devised, so such a system must allow for supplementary notes that modify or add completeness to a description.

As this committee was trying to devise a method and blanks for recording facts about a pupil in abbreviated form, it was necessary to agree upon working objectives for producing the kind of forms that would serve the purposes desired. The following objectives were used.

#### WORKING OBJECTIVES FOR RECORDS AND REPORTS

1. Any form devised should be based on the objectives of teachers and schools so that a continuing study of a pupil by its use will throw light on his successive stages of development in powers or characteristics believed to be important.
2. The forms dealing with personal characteristics should be descriptive rather than of the nature of a scale. Therefore "marks" of any kind, or placement, as on a straight line representing a scale from highest to lowest, should not be used.
3. Every effort should be made to reach agreement about the meaning of trait names used, and to make their significance in terms of the behavior of a pupil understood by those reading the record.
4. Wherever possible a characterization of a person should be by description of typical behavior rather than by a word or phrase that could have widely different meanings to different people.
5. The forms should be flexible enough to allow choice of headings under which studies of pupils can be made, thus allowing a school, department, or teacher to use the objectives considered important in the particular situation, or for the particular pupil.
6. Characteristics studied should be such that teachers will



be likely to have opportunities to observe behavior that gives evidence about them. It is not expected, however, that all teachers will have evidence about all characteristics.

7. Forms should be so devised and related that any school will be likely to be able to use them without an overwhelming addition to the work of teachers or secretaries.

8. Characteristics studied should be regarded not as independent entities but rather as facets of behavior shown by a living human being in his relations with his environment.

This last objective is a fundamental one. It has been observed in the work on both evaluation and recording, and must be kept in mind in considering whatever has been produced. The one great danger in the use of any forms that offer opportunity for recording facts about people is that those who use them may revert to the idea of "marking," using the material on the forms as a scale for rating, instead of as an abbreviated basis for description of the person's behavior in some area or under some conditions. The various record forms too should be considered as supplementing each other so as to give a more complete description of the individual than a single form could present.

It should be emphasized that no form produced in this study is believed to be final, or to be the only kind of form for its purpose. Because of the generosity of the contributing foundations and the willingness of the committee members to give their time and effort, a more extensive and intensive study of recording has been made than had been possible before. There is reason to hope, therefore, that these forms may prove suitable for many institutions, particularly in view of their wide flexibility. For other institutions they may need modification, for still others they may prove suggestive in detail or principles. In any case the committees concerned hope the objectives and the material developed will prove worthy of study and trial, though the members

are far from being dogmatic about the form or content of what is now offered.

While that which has been done by these committees represents the most organized work accomplished in recording and reporting, since it involves the cooperation of those in many colleges and schools, the achievements of various of the cooperating schools working individually in devising forms to fit their own particular needs also deserves mention. Committees of faculty members studied the conditions and needs of their institutions and arrived at interesting and valuable methods of collecting and recording information about their pupils.

It is obviously impossible to reproduce and discuss the forms produced by such efforts, but other schools may profit by consultation with cooperating schools whose problems seem similar to their own.

## Chapter X

### BEHAVIOR DESCRIPTION

~~~~~  
Much of the foregoing philosophy was developed while the Committee on Records and Reports¹ was making a preliminary study of its first record-making assignment, which concerned the study of personal characteristics. This study began with exploration of what had previously been done in this field. The committee found many attempts to clarify and organize the study of human beings, with little agreement on the terms used or the methods employed. From the great number of people-describing words in the language, however, certain ones had attained somewhat common usage. The first survey of terms used by various agencies to describe people produced over 150 terms, and a later study made by Dr. Rothney listed over 260 trait names.

All of these words were considered and compared. It was found that they fell into sets, each set composed of words having somewhat the same meaning, so that the number of markedly distinct characteristics was only a fraction of the number of names of traits. Each set was considered by itself

¹ COMMITTEE ON BEHAVIOR DESCRIPTION. (Members and those added during the work. Institutional affiliations are those for the time of appointment.) Miss Helen M. Atkinson, Horace Mann School for Girls; E. Gordon Bill, Dartmouth College; Carl Brigham, Princeton University; Oscar K. Buros, Research Assistant 1933-35, Rutgers College; Mrs. Cecile Fleming, Horace Mann School; Mrs. Anne Rose Hawkes, The Carnegie Foundation; Miss Frances Knapp (deceased), Wellesley College; Robert D. Leigh, Bennington College; William S. Learned, The Carnegie Foundation; John Lester, The Hill School; Rollo Reynolds, Horace Mann School for Girls; Eugene R. Smith, Chairman, The Beaver Country Day School; John Tildsley, Associate Superintendent of Schools, New York City; Ben Wood, Co-operative Test Service; Stanley R. Yarnall, Germantown Friends School; John W. M. Rothney, Research Assistant, Secretary, Harvard University.

until the committee members agreed on the term or terms that best expressed its fundamental meaning. From the resulting list of key words the first group of characteristics was chosen for further study.

The criteria for choosing the characteristics to be used were:

1. *Importance.* The ones chosen should be worth observing because they throw light on the person being studied.

2. *Observability.* They should be such that some at least of a pupil's teachers will have opportunity to observe significant behavior in relation to them.

3. *Completeness.* Taken together they should give a reasonably complete picture of the person as seen by the adults dealing with him.

4. *Differentness.* They should be sufficiently independent so that teachers can distinguish between them and so that intercorrelations will not be too high.

From the beginning, the members of the committee were agreed that the evidence from research did not justify a method of rating, or any type of scale for judging personal characteristics, such, for example, as one constructed along a straight line, or one composed of named points with supposedly equal intervals between them. It questioned the use of undefined terms for designating degrees of excellence or lack of it, and believed that it was unlikely that intervals on a line or other scale had any accuracy in terms of their relative size or importance.

Furthermore the committee was not much interested in a scale even if it could have been constructed. It hoped rather for something that would encourage and help teachers to observe and analyze behavior and from the evidence obtained to reach a better understanding of their pupils as living functioning human beings.

The members, as has been shown, were definite in their desire to eliminate comparisons except as they were implicit

in any descriptive material. They therefore set as their goal a form that:

1. at the time of a single use of it, would, through descriptions of behavior, present a picture of a person not only in terms of his commonest (modal) behavior, but also in terms of the range and variety of his behavior under different conditions;
2. over a period of time would, through a series of studies and recordings, constitute a record of development in significant characteristics.

It would be difficult for any one who had not worked on such an undertaking to realize the difficulties encountered. The members of the committee covered a wide range of experience and specialization that naturally influenced their ideas of the work of the committee and their conceptions of the use and meanings of certain terms. Some, at the beginning, were even skeptical about the possibility of working out anything of value. Frequently hours had to be spent in the discussion of the meanings of a few words whose use seemed necessary. Little by little, however, techniques of work developed and language difficulties became fewer. The final form includes only material on which the committee reached substantial agreement.

The committee's first achievement was the choice of a group of characteristics and the development of blanks for recording behavior in terms of them. A manual for teachers was also written. The cooperating schools were asked to study groups of pupils by the use of the forms and manual, to send the completed blanks to the committee, and to make suggestions for revisions. Mr. Oscar K. Buros of the research staff and an assistant studied the results and from their analysis made suggestions for changes. A blank and a manual incorporating the revisions decided on was next prepared. After rather a large number of pupils had been studied

the intercorrelations among descriptions were worked out at Columbia University under the direction of Dr. Ben Wood. The figures showed that either some names of traits conveyed too nearly the same meaning to teachers despite the committee's attempt to differentiate their meanings, or else certain characteristics were so closely related that they tended to appear in similar ways in many situations. In either case the aims for the undertaking were not being achieved.

The committee made the changes indicated. It also added to the scope of the information asked for, since some valuable facts seemed to be omitted, and rewrote and enlarged the manual. Further trial, experimentation and testing resulted in still further changes, though with less radical alterations in later steps. Eventually a considerable body of material was again submitted to correlation study, this time at Harvard University by Dr. Rothney. This study found that the characteristics were sufficiently different and the judgments of the teachers sufficiently well made so that the reports were significant descriptions of the pupils.

Even after this the committee again called for criticisms and suggestions from the schools and tried to refine its work. It is hoped that it is now in such form that it will have value to schools in general, and perhaps to more advanced institutions.²

It will be clear from the material itself that the method of studying pupils devised by the committee depends on the supplying of descriptions of the different kinds of behavior that are likely to be observed in relation to the characteristics chosen. The descriptions made by the committee

² The form, modified in its text material only by the addition of two characteristics and two additional questions was used by Dartmouth College in "The Dartmouth Visual Survey of The Dartmouth Eye Institute." It is said to have served the purpose of the study successfully. The cards have also been used to study the students in a college dormitory. It is likely that a form planned especially for college use will eventually be published.

are designed to define what might be called types or classifications of behavior in terms of each characteristic. The use of carefully worded standard definitions in place of teachers' own wordings is intended to bring about a more nearly common understanding of the characteristics themselves and of the persons described. The form resulting also decreases greatly the time required for recording and for using the record for purposes of interview or transfer.

In general, all teachers having opportunity to know a pupil would be expected to describe him by the use of this material. The combined reports, which would appear on the Behavior Description card, would show the pupil's most common behavior, as well as the range of behavior under different conditions.

It is recommended that the descriptions be recorded twice a year through the six years of junior and senior high school. To the degree that the information covers this period the card becomes a record not only of what a pupil is like at any one time but of his many-sided development through this period of his growth.

USE OF RECORD CARDS

To show the manner in which the classifications are considered and used in school practice the entire section on "Creativeness and Imagination" is quoted here from the manual.

CREATIVENESS AND IMAGINATION

NOTE: The question whether what is created has been created before does not enter into this discussion. Newness to the person in question, and the extent of the contribution he himself makes, determine the amount of creativeness shown. Creation includes not only originating entirely, but also recombining old elements and seeing new relationships. Some characteristics that tend toward creativeness are:

the desire and habit of trying new things, of putting things together in new combinations (experimentation),

the ability to think new things, an art form, a melody, a new concept, a new situation (imagination),

the ability to organize, direct or control new combinations of people or things (executive manipulation).

TYPE IA. General: those who approach whatever they do with active imagination and originality, so that they contribute something that is their own.

TYPE IB. Specific: those who make distinctly original and significant contributions in one or more fields.

Discussion: For secondary school pupils this might occur in writing, the fine or applied arts, music, drama, or research in scientific or other fields.

Examples: One may show the possession of this trait by:

1. Expressing one's emotions and thoughts through such media as language, arts and crafts, music, or drama. This might result in the writing of poems, stories or essays, in the conception and execution of pictures, statues, costumes, or stage sets, or in one or more of various other such expressions.
2. So expressing an old idea that it is reinterpreted through a new viewpoint or a different organization of material.
3. Using logical processes with such imagination that he sees implications and relationships that open new fields of thought or throw light on old ones.
4. Bringing to the planning and activities of the day thinking and action that result in improved procedures. This might appear in the formulation and carrying out of a procedure for study investigation, the accomplishment of a task, or the manipulation of a group.
5. So completely projecting oneself into a situation that it becomes his own. One can listen creatively to a symphony, or can interpret with originality the one whose part he plays in dramatization.

6. Combining elements (as in an invention) to produce a new result or improve a procedure.

TYPE II. Promising: those who show a degree of creativeness that indicates the likelihood of valuable original contribution in some field, although the contributions already made have not proved to be particularly significant.

Discussion: This includes those who show imagination and approach their problems creatively, although—perhaps because of lack of experience or of opportunity in the fields in which they will eventually contribute—they have as yet shown indications rather than demonstrated accomplishments.

TYPE III. Limited: those whose general attitude shows the desire to contribute their own thinking and expression to situations, but whose degree of imagination and originality is not in general high enough to have much influence on their accomplishments.

Discussion: A person of this type may make occasional contributions of some general value where particular experience or other favorable influences make this possible, or may from time to time show originality in details rather than in general situations.

TYPE IV. Imitative: those who, while they make little or no creative contributions themselves, yet show sufficient imagination to see the implications in the creation of others and to make use of their ideas or accomplishments.

TYPE V. Unimaginative: those who have given practically no evidence of originality or creativeness in imagination or action.

The "Type" numbers are used for convenience in referring to the descriptions of different kinds of behavior, and the words "General," "Specific," and so forth are key words defining each type of behavior as well as one or two words could be found to do it.

Under some characteristics two types keep the same number but with letters added, as in IA and IB in this case. This occurs where the committee wishes to indicate two related types of behavior that differ only in the way in which the individual uses, or is limited in the use of, the character-

istic in question. Both IA and IB in this example indicate a highly creative approach to problems on the part of those they describe, but of two listed under these definitions the one under IA might be thought of as applying his creative ability more extensively, while the one described by IB would respond less generally, but quite possibly with equal or greater intensity, to the particular stimuli that do arouse his creativeness.

The Behavior Description card, because of its size, which is that of a filing envelope for an 8½" by 11" file, cannot easily be shown in this volume. It is possible however to describe what is most significant about it. It consists of:

1. A listing of characteristics and the descriptive classifications under them.
2. Spaces opposite the classifications that make it possible to include on the card the study of a pupil over the six years of junior high school and senior high school, or over the seventh and eighth grades and the four year secondary school.
3. A key system for use in recording the judgments of teachers. This will be illustrated later under "Responsibility-Dependability."
4. A considerable space for "General Comment."

The entire list of characteristics that use defined descriptions of types of behavior follows as it appears on the filing card.

RESPONSIBILITY-DEPENDABILITY

Type

- | | |
|--|---|
| RESPONSIBLE AND RESOURCEFUL: Carries through whatever is undertaken, and also shows initiative and versatility in accomplishing and enlarging upon undertakings. | 1 |
| CONSCIENTIOUS: Completes without external compulsion whatever is assigned but is unlikely to enlarge the scope of assignments. | 2 |

	Type
GENERALLY DEPENDABLE: Usually carries through undertakings, self-assumed or assigned by others, requiring only occasional reminder or compulsion.	3A
SELECTIVELY DEPENDABLE: Shows high persistence in undertakings in which there is particular interest, but is less likely to carry through other assignments.	3B
UNRELIABLE: Can be relied upon to complete undertakings only when they are of moderate duration or difficulty and then only with much prodding and supervision.	4
IRRESPONSIBLE: Cannot be relied upon to complete any undertaking even when constantly prodded and guided.	5

CREATIVENESS AND IMAGINATION

GENERAL: Approaches whatever he does with active imagination and originality, so that he contributes something that is his own.	1A
SPECIFIC: Makes distinctly original and significant contributions in one or more fields.	1B
PROMISING: Shows a degree of creativeness that indicates the likelihood of valuable original contribution in some field, although the contributions already made have not proved to be particularly significant.	2
LIMITED: Shows the desire to contribute his own thinking and expression to situations, but his degree of imagination and originality is not in general high enough to have much influence on his accomplishments.	3
IMITATIVE: Makes little or no creative contributions, yet shows sufficient imagination to see the implications in the creation of others and to make use of their ideas or accomplishments.	4
UNIMAGINATIVE: Has given practically no evidence of originality or creativeness in imagination or action.	5

INFLUENCE

CONTROLLING: His influence habitually shapes the opinions, activities, or ideals of his associates.	1
CONTRIBUTING INFLUENCE: His influence, while not control-	

Type

- ing, strongly affects the opinions, activities, or ideals of his associates. 2
- VARYING: His influence varies, having force when particular ability, skill, experience, or circumstance gives it opportunity or value. 3
- COOPERATING: Has no very definite influence on his associates, but contributes to group thinking and action because of some discrimination in regard to ideas and leaders. 4
- PASSIVE: Has no definite influence on his associates, being carried along by the nearest or strongest influence. 5

INQUIRING MIND

- GENERAL: Responds with consistent, active, and deep interest to any intellectual stimulus and uses to good advantage various sources of information. 1
- SPECIFIC: Responds with consistent, active, and deep interest only to stimuli arising in specific fields or problems. Uses effectively the sources available for such purposes. 2
- LIMITED: Somewhat sensitive to stimuli arising from limited fields, but engages in exploration and investigation only when a general plan of attacking the problem is indicated to him. 3
- DIRECTED: Responds to stimuli in a limited field of interests but is impelled to act only when both the plan and the details of procedure are definitely outlined for him. 4
- UNRESPONSIVE: Rarely seems to be sensitive to any intellectual stimulus and shows little or no ability to use the tools and methodology of exploration and investigation. 5

OPEN-MINDEDNESS

- DISCRIMINATING: Welcomes new ideas but habitually suspends judgment until all the available evidence is obtained. 1
- TOLERANT: Does not readily appreciate or respond to opposing viewpoints and new ideas, although he is tolerant of them and consciously tries to suspend judgment regarding them. 2

- PASSIVE: Tolerance of the new or different is passive, arising from lack of interest or conviction. Welcomes, or is indifferent to, change, because of lack of understanding or appreciation of the new or of that which it replaces. 3
- RIGID: Preconceived ideas and prejudices so govern his thinking that he usually ends a discussion or an investigation without change of opinion. 4
- INTOLERANT: Is actively intolerant; resents any interference with his habitual beliefs, ideas, and procedures. 5

THE POWER AND HABIT OF ANALYSIS; THE HABIT OF
REACHING CONCLUSIONS ON THE BASIS OF
VALID EVIDENCE

- HIGHLY ANALYTICAL: Habitually makes an analytical approach to his problems, assembling the facts, showing a clear perception of their relationships and implications, and thinking through the situation to well founded conclusions. 1
- INCOMPLETE: Makes an intelligently analytical approach to his problems but is more limited in ability to assemble the facts completely, and to see their relationships or their implications. 2A
- IRREGULAR: On occasion shows unusual analytical power but does not do so habitually. 2B
- UNDEVELOPED: Shows signs of analytical power, but because of fears, the domination of others, or some other inhibiting agency, has not yet developed it to any high degree. 3A
- LIMITED: Is able to pursue reasoning processes if aided by some guidance and direction. 3B
- PASSIVE: His approach to a problem is not an analytical one, though he may be able to appreciate a train of reasoning or to follow one laid out by some one else. 4
- UNREASONING: Seems unable to analyze even a fairly simple situation, tending rather to rely on memory as a substitute for logic. Accepts statements and results without attempting to reason about them. 5

SOCIAL CONCERN

Type

- GENERALLY CONCERNED: Shows an altruistic and general social concern and interprets this in action to the extent of his abilities and opportunities. 1
- SELECTIVELY CONCERNED: Shows concern by attitude and action about certain social conditions but seems unable to appreciate the importance of other such problems. 2
- PERSONAL: Is not strongly concerned about the welfare of others and responds to social problems only when he recognizes some intimate personal relationship to the problem or group in question. 3
- INACTIVE: Seems aware of social problems, and may profess concern about them, but does nothing. 4
- UNCONCERNED: Does not show any genuine concern for the common good. 5

EMOTIONAL RESPONSIVENESS

- TO IDEAS: Is emotionally stirred by becoming aware of challenging ideas. 1
- TO DIFFICULTY: Responds emotionally to a situation or problem challenging to him because of the possibility of overcoming difficulties. 2
- TO IDEALS: Responds emotionally to what is characterized primarily by its personal or social idealism. 3
- TO BEAUTY: Responds emotionally to beauty as found in nature and the arts. 4
- TO ORDER: Responds emotionally to perfection of functioning as it is seen in organization, mechanical operation, or logical completeness. 5

SERIOUS PURPOSE

- PURPOSEFUL: Has definite purpose and plans and carries through to the best of his ability undertakings consistent with this purpose. 1
- LIMITED: Makes plans and shows determination in attacking short-time projects that interest him, but has not yet thought out goals for himself. 2

Type

- POTENTIAL: Takes things as they come, meeting situations somewhat on the spur of the moment, yet may be capable of serious purpose if once aroused. 3
- UNRELIABLE: Makes plans that are fairly definite, but cannot be counted on for the determination to carry them through. 4
- VAGUE: Is likely to drift without the decision and persistence that will enable him to carry out his vaguely conceived plans. 5

SOCIAL ADJUSTABILITY

- SECURE: Appears to feel secure in his social relationships and is accepted by the groups of which he is a part. 1
- UNCERTAIN: Appears to have some anxiety about his social relationships although he is accepted by the groups of which he is a part. 2
- NEUTRAL: Shows the desire to have an established place in the group, but is, in general, treated with indifference. 3
- WITHDRAWN: Withdraws from others to an extent that prevents his being a fully accepted member of his groups. 4
- NOT ACCEPTED: Has characteristics of person or behavior that prevent his being an accepted member of his group. 5

WORK HABITS

- HIGHLY EFFECTIVE: A pupil having highly effective work habits would be likely to reach the maximum accomplishment for one of his ability. 1
- ADEQUATE: A pupil having adequate work habits would accomplish all that would commonly be expected of one of his ability. 2
- PROMISING: While his habits are not yet adequate, they show promise of becoming so. 3
- LIMITED: Has work habits that are adequate only for simple situations, or are limited by the lack of development of some elements that make for efficiency. 4
- INEFFECTIVE: Has not developed his work habits to the point where he can work efficiently. 5

It will be seen that the subheads under "Emotional Responsiveness" are not exclusive, since a pupil might respond to any number of them. In this respect the treatment of this characteristic differs from that of the others.

The key for recording teachers' judgments, which a school can extend as it seems necessary, lists abbreviations that show the type of opportunity a teacher has had for observing the pupil being described.

The following example will show how this is used.

Under "Responsibility-Dependability" six types of behavior are defined. They will be listed by their numbers and key words, and the judgments of nine teachers about a pupil will be shown as they would appear on a filing card:

1	Responsible	M—HR
2	Conscientious	N.S.—S.S.—E.—F.
3A	Generally Dependable	A—Mu
3B	Selectively Dependable	
4	Unreliable	P
5	Irresponsible	

This indicates that the teacher of mathematics and the home-room teacher believe the boy fits the definition of Type 1, that teachers of natural science, social science, English, and French place him as Type 2, that art and music teachers would describe his behavior as of Type 3A, while the one in charge of physical education would place him under Type 4.

The total picture of this boy's behavior (but only in respect to his responsibilities) shows him to be highly conscientious in meeting the demands of academic work and of the group (home-room) with which he is closely connected. It also shows that for some reason he is not so highly dependable in the arts, and that he is failing to meet with any consistency the obligations that are related to physical education. It is not, of course, safe to make positive judg-

ments about the arts and the physical education from this information alone. Evidence about the other characteristics may throw light on what is shown here, and personal relationships, home obligations, or other factors may enter into the situation.

It is evident from this example that a principal, supervisor, or guidance officer can not only obtain information from the numerical distribution of judgments and the situations in which extremes of behavior occur, but also can take into account what he knows about teachers and courses, in this way reaching a more accurate understanding of the pupil than would otherwise be possible.

While one outside an institution cannot obtain so complete an understanding as this, information from this card and the comment of a supervisor, recorded on such a form as that used for transfer to college (Chapter XII), can give a very accurate description for the use of a college admissions officer or a prospective employer.

The fact that the classifications under any heading on the card were not intended to constitute a rating scale, cannot be too strongly emphasized. The committee was also agreed in the belief that the classifications obtained could not even be said to define orders of excellence, since there was no certainty that some earlier classes were better than others that were later in the lists, nor that behavior of a certain type was best for all kinds of people under all kinds of conditions. It is true that the first classifications generally describe behavior that would be considered highly desirable, that the last are, in general, not indicative of such favorable traits, and that there is a general decrease of desirability through most of the classes. It cannot be assumed, however, that each class is below the preceding one or above the following one in desirability. Neither can it be taken for granted that where there is evident decrease in desirability the intervals are equal, or in any fixed relationship to one another.

The classifications are therefore simply items of the description of a person in terms of his behavior under various conditions, as judged by a number of practiced and supposedly impartial observers. It is of course true that the limited number of descriptions cannot exactly describe all possible kinds of behavior. It is believed, however, that the definitions will usually fit closely enough for practical purposes, particularly since when necessary they can be modified by further comment.

In addition to the characteristics so far listed there are four on the card about which the only judgment asked for each is whether it is present or absent to a marked degree. The four, which are defined on the blank, are PHYSICAL ENERGY, ASSURANCE, SELF RELIANCE, and EMOTIONAL CONTROL.

Two other details are worthy of notice. At the end of the printed material there is a place for indicating the judgment of the faculty in regard to the success of the pupil in four broad fields of thought and activity. These are "abstract ideas and symbols," "people," "planning and management," and "things and manipulation." It is thought that where there are marked differences in success in these areas the evidence may prove valuable in guiding a pupil toward suitable after-school experiences. The information may help to decide whether or not the pupil should go to college, and if so to what kind of a college, whether or not he should undertake some form of specialization, what kind of a job he should try to obtain.

The other detail is the large space left for "comment." This is useful for the recording of information that explains, amplifies or brings into relationship the description on other parts of the card.

Successful use of the behavior description material requires study of the manual and careful following of its directions. At first this may seem to require more time than a teacher is able to give. However, the time needed for re-

ording will grow rapidly less as one becomes familiar with the method used, particularly if a teacher is already observing and analyzing the behavior of his students to the extent any good teacher should. It is the conviction of the committee that time spent in better understanding of a pupil does, in any case, justify itself in better relationships and more effective work.

It is interesting to know in this connection that one public school system has adopted this form for the study of 12,000 pupils in junior and senior high school and expects soon to extend it to another 6,000 pupils. Some colleges, as has been said, have found the card valuable in obtaining and recording facts about behavior, and many types of schools are experimenting with the material. Samples have gone to other countries, even to Russia and South Africa, as well as to most sections of the United States.

SUMMARY OF ADVANTAGES

This form replaces "rating" as a basis for studying individuals by *description of behavior* as observed by adults having a variety of associations with the one studied.

In general it shows, for any characteristic, a pupil's *most common behavior and range of behavior*. Where no mode appears, the judgments being so scattered as to have no modal point, that fact in itself has significance, the particular implications depending on the pattern of judgments and the characteristics in question.

Taken as a whole, the card when filled in gives a reasonably *complete picture* of the person's behavior because the characteristics, each of which emphasizes one facet of behavior, combine to form quite a comprehensive description of him.

The material is in such form that it can very *quickly be transferred* to a cumulative record card or a college entrance blank, or be used as a basis for an interview with parents.

On a college entrance blank the information can show the pupil's most common behavior and the number of reporters who observed it, and can indicate the range and under what conditions extremes occur. The form in Chapter XII shows such a transfer from this card.

Chapter XI

TEACHERS' REPORTS AND REPORTS TO THE HOME

During the Study various schools wrote to the chairman of the Committee on Evaluation and Recording asking about tendencies in reports to parents and expressing dissatisfaction with existing forms. A sub-committee¹ was therefore appointed to investigate the practices of schools, to analyze tendencies in reporting, and to make recommendations of forms for teachers' use and for sending reports to the home. This committee's first step was to collect report forms from schools of various kinds, and to ask the schools to say how and why present practices were unsatisfactory and to comment on what reports should be. The report cards obtained were carefully studied, and the criticisms and suggestions sent in by the schools were analyzed. Quite a number of schools, however, sent no forms, saying that they had nothing that would be of any help in the undertaking. It became clear at once that the most general demand was for something that would replace numerical or letter marks, and would give more usable information about a pupil's strengths and weaknesses.

Many schools were convinced that the single mark in a subject hid the facts instead of showing them clearly. The mark was, in effect, an average of judgments about various elements in a pupil's progress that lost their meaning and

¹ The members of the committee were: Helen M. Atkinson, Derwood Baker, Genevieve Coy, Rosamond Cross, Burton P. Fowler, I. R. Kraybill, Elvina Lucke, Eugene R. Smith, Chairman, John W. M. Rothney, Research Assistant.

their value when thus combined. The schools believed that the value of a judgment concerning the work done by a pupil in any school course or activity depended on the degree to which that judgment was expressed in a form that showed his strengths and his weaknesses and therefore presented an analyzed picture of his achievement that would be a safe basis for guidance.

There was also a feeling that marks had become competitive to a degree that was harmful to both the less able and the more able, and that they were increasingly directing the attention of pupils, parents, and even teachers, away from the real purposes of education toward the symbols that represented success but did not emphasize its elements or its meaning.

The commonest method of replacing marks proved to be that of writing paragraphs analyzing a pupil's growth as seen by each teacher. This method is an excellent one, since good descriptions by a number of teachers combine to give a reasonably complete picture of development in relation to the objectives discussed. On the other hand, a report in this form is very time-consuming for teachers and office, as well as difficult to summarize in form for use in transfer and guidance. The committee decided on a compromise that would make place for giving definite information about important objectives in an abbreviated form and would allow for supplementing this with written material needed to modify or complete the information.

To find the objectives, the list collected by the Evaluation Staff and the forms worked out by the committees for the various subject fields (Chapter XIII) were studied. It was discovered that there were five objectives that were common to all fields and experiences, and about which knowledge would be particularly valuable to parents as well as to pupils. These five objectives were therefore chosen as headings to be reported on by all teachers and to be used

in reports to the home. The wording adopted for them is not, however, identical with the wordings on the forms used in subject fields. The reason is that this committee had to draw from the large amount of information asked for on the subject forms that which could be condensed into simple phrases that would have meaning and importance on a report to the home. The headings follow:

Success in Achieving the Specific Purposes of the Course
Progress in Learning How to Think

Effectiveness in Communicating Ideas:

Oral

Written

Active Concern for the Welfare of the Group

General Habits of Work

The question of classifications to indicate degrees of success or growth in relation to these objectives proved a difficult one. After much discussion and experimentation it was decided to take as a point of departure the usual expectation for one of the age group and the background of the pupil in question. Two classifications above and two below are used. They are defined as follows:

IS OUTSTANDING: The pupil has reached an *outstanding* stage of development in the characteristic and field indicated: that is, a stage distinctly above that usual for pupils of the same age and similar opportunities.

IS ABOVE USUAL: The pupil has reached a stage of development somewhat *higher* than usual, perhaps with promise of eventually reaching a superior level.

IS AT USUAL STAGE: The pupil is at approximately the *usual* stage of development for age and opportunity.

IS BELOW USUAL: The pupil is sufficiently *below* the usual stage in this field to need particular help from the home and school or greater effort on the part of the pupil.

IS SERIOUSLY BELOW: The pupil is *seriously below* an acceptable standard in the field indicated.

In this particular these forms depart somewhat from the descriptive method that is emphasized in the work of all the committees, though taken as a whole these blanks are still highly descriptive. This departure, however, should not be thought of as too inconsistent, since the purpose of these forms affected to some extent the method to be used. It seems likely that the time will come when each pupil is judged primarily in accordance with his ability and his opportunities, rather than in comparison with others. There is still demand, however, for information that will tell parents with some definiteness where their children are showing strengths or weaknesses as judged by normal expectations. These forms try to meet that demand and at the same time to describe the pupil's progress in a way analytical enough to give helpful guidance.

In addition to the section that tells the degree of success a pupil is achieving in the five objectives listed, there are three other sections of the report. The first gives opportunity for the teachers to point out weaknesses a pupil should particularly try to eradicate. There are eight of these listed, and the subjects in which the weaknesses are evident are shown on the home report:

- Accuracy in following directions
- Efficient use of time and energy
- Neatness and orderliness
- Self-reliance
- Persistence in completing work
- Thoughtful participation in discussion
- Conscientiousness of effort
- Reading

There is also opportunity for the teachers to report on the pupils' likelihood of success in continuing to work in their fields, both in later years in school and in advanced institutions.

A section for "General Comment" appears on the teacher's report, and on the report to the home. Some schools copy the most valuable of the teachers' comments upon the home report form. Others summarize criticisms and suggestions in this space. Occasionally so much of value should be sent that an attached sheet must be used, but in general the space for comment seems to be sufficient.

In all the details that have been mentioned the teachers' report and the home report are identical, although they differ in arrangement, since the home report is designed to combine the reports of all the teachers into a single form that can be read easily.

There are two forms of the report to the home. They include the same material but differ in arrangement in a way that produces somewhat different emphases. Form A tends to emphasize the objectives in which a pupil is strong or weak, while Form B goes further in showing a pupil's degree of success in individual subjects. A school can choose either form or can do as a school represented on the committee has done. This school liked the completeness of the teachers' reports so well that it decided to send copies of all of them to the parents instead of using the combined report form.

While one of the greatest values of these forms is the way in which they provide for guidance by analyzing a student's progress instead of trying to express several factors in one "mark," the form has other advantages.

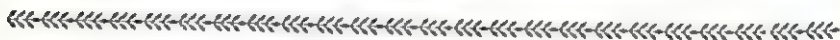
An important one is the degree to which it directs the minds of pupils, parents, and teachers away from marks toward the fundamental objectives with which pupils should be concerned. Incidentally, in this procedure it is not easy to compare two reports in a way to make the less able pupil feel inferior or the more able one become smug, for in such an analysis even the poorest student is likely to find some appreciation, while the best student is likely to discover some weaknesses to be corrected.

It hardly seems necessary to point out the fact that this form, like the "Behavior Description," attempts to describe somewhat fully a phase of the behavior of a person. In this case, it is principally the pupil as one who is learning and developing mental power that is observed. As in the other form, the pupil is studied by a number of teachers, and the mode and distribution of response in different environments is recorded. The comment appearing on the form sent to the parents becomes an analysis of what is shown under the various headings, and a recommendation of ways in which the pupil can be helped to overcome his weaknesses and use his ability more effectively.

A word of warning about the introduction of such report forms may not be amiss. Pupils and parents should receive some explanation of the meaning of the information given so that they will not be confused by the very completeness of what is said and will not be antagonized by the unfamiliar material.

Chapter XII

FORM FOR TRANSFER FROM SCHOOL TO COLLEGE



CONFIDENTIAL REPORT TO THE COMMITTEE ON ADMISSION

The need for a new transfer form has been widely recognized. Schools everywhere wish a uniform blank, since the present waste of the time of school officers, because of the wide variety of forms used by different colleges, has reached serious proportions.

Recognition of the extent to which marks and "units" are preventing schools and colleges from giving their best service to individual students, and are interfering with educational progress, also becomes daily more widespread. The reasons for replacing marks by analyses were discussed in relation to reports to the home. Units, too, become the objectives for which pupils strive, sometimes with little consideration of the methods by which they are obtained. In many schools, also, reorganized courses, activity programs, and long time researches (though on a secondary school level) have so changed the schedule that the definition of a unit no longer has meaning.¹ A college entrance form with less emphasis on marks and units can help greatly toward overcoming the abuses that are of so much concern to the schools. Then, too, it is increasingly recognized that education should have a degree of continuity that has not yet existed, and that information useful for guidance should be

¹ The Carnegie Foundation for the Advancement of Teaching has been considered responsible for the adoption of units as a measure of work accomplished in school. Various officers of the Foundation have now, in speeches and writing, said that units no longer have value.

provided by the schools for use in college. The entrance blank seems a natural place for such information.

As an example of this general movement, the Committee on School and College Relations of the Educational Records Bureau, which is composed of school and college representatives, has sent bulletins to the colleges emphasizing needed changes in information required at entrance, and has published² the answers of the colleges, which show quite general willingness to cooperate in making the changes. Another bulletin has recently been sent to the colleges, and the answers will soon be published. A striking example of the interest taken by educators in the various needs being discussed is the fact that the Educational Records Bureau Committee has given the Committees on Records and Reports of the Progressive Education Association standing as sub-committees of its own in order to keep in touch with their work, and to lend its support to whatever promises progress in better school and college relations.

This dissatisfaction with entrance blanks was focussed by the necessity, under the Eight-Year Plan, of developing an entrance form that would accomplish two objectives:

1. Have such a range of flexibility and such carefully chosen items that it would not restrict any school's curriculum or methods.
2. Provide for information complete enough to replace effectively the data that was omitted under the special plan for the cooperating schools, and significant enough to assist in the guidance programs of the colleges.

The Committee on Evaluation and Recording appointed a sub-committee³ to work on this problem. This committee,

² Published by the Educational Records Bureau, 437 West 59th Street, New York City.

³ The members of this committee were: Victor L. Butterfield, Genevieve L. Coy, Albert B. Crawford, Ruth W. Crawford, Burton P. Fowler, Elvina Lucke, Herbert W. Smith, Eugene R. Smith, Chairman, Arthur E. Traxler, John W. M. Rethney.

after studying previous reports on the subject, explored the forms in use, especially those prepared by groups of colleges. All forms that had wide use were analyzed, and their items were listed with ratings of their prevalence in present blanks.

The committee also asked schools for their criticisms of entrance blanks and their suggestions for improvement, and on the basis of the two surveys a new blank was devised and has been in use by the cooperating schools with the very large number of colleges to which they send students.

The first page of the form⁴ is given over very largely to a tabular history of the courses the pupil has taken in school, and a combined recommendation and prediction for work in college. This table allows a school that wishes to do so to record only traditional marks and units, but it also allows for courses not easily expressed in units and not recorded by marks, since it has space for final recommendations in the major departments most likely to be presented for entrance or followed in college, and provides blank spaces for additions. If this form were being prepared now it would probably have no column for units, but when it was being devised the movement for omission of unit equivalents in Statements of Credit had not reached the point it has since attained.

The second page is given to test records and includes a blank space for "Summary Interpretation" of tests whose results are not easily expressed in numerical forms. Such tests include ones described in the "Evaluation" section of this report, as well as tests of primary abilities and others that have important sub-heads.

The particular contribution of the third page is the tabular form for the description of a pupil's behavior, and a resulting characterization of him. The table is based on definitions of the characteristics and the sub-heads under them as they are given in the "Manual of Behavior Description," and is supposed to be used with those definitions. (See Chap. X.)

⁴ The form is between pp. 469-497.

NOTES

SUMMARY INTERPRETATION OF OTHER NEW-TYPE TESTS

(Enter here any results requiring analysis in words.)

PRINCIPAL'S RECOMMENDATION

John A. Doe

STUDENT'S NAME _____

IS ENTITLED TO HONORABLE DISMISSAL FROM THE SCHOOL.

GRADUATION IS EXPECTED IN June, 1941

MY SUMMARIZED RECOMMENDATION OF HIM FOLLOWS:

We recommend John Doe as an acceptable member of the student body and as a fully prepared and capable candidate for college.

April 15, 1947

Walter E. Draft. Principal

SIGNATURE

The method of recording, which reports the judgments of all the teachers dealing with a pupil, gives two very important facts about his behavior in respect to any one of the characteristics:

1. His most common type of behavior.
2. The range of behavior on one or both sides of the modal heading.

For example:

WORK HABITS	Highly effective English	Adequate	Promising M-5	Ineffective	Limited Math. Sci.
----------------	-----------------------------	----------	------------------	-------------	-----------------------

This would indicate:

a. that the pupil's work habits had been judged by eight people, of whom five thought they accorded best with the definition of "Promising";

b. that in English, because of response to the subject, the influence of the teacher, or some other reason, his habits seemed "Highly Effective";

c. that in mathematics and science his habits were as defined under "Limited."

These facts might have great significance both for consideration of a candidate for college, and for guidance if he was accepted.

A school that did not wish to use any tabular method of description might omit the use of this table and describe the candidate in paragraph form on the next page.

The fourth page is left for the school's comments. It may replace the table on page three but in any case it gives the opportunity to supplement, modify, and summarize the rest of the blank. It ends with a place for the definite recommendation of the school head, an item that all colleges seem to value.

Other items on the blank are self-explanatory and differ only slightly from commonly used headings.

All the items most commonly asked for by the colleges, and

possible for the schools to furnish, are included on the blank, while those that have been found to have little importance in actual use have been omitted. An occasional college asks for one or two additional facts, which can usually be given under "Comment" if no other place seems more suitable for them.

This form has been in successful use for four years, and its use is spreading to schools outside of the Study, sometimes through initiation by a school, sometimes through its adoption by a college. It is hoped that in its present, or a modified, form it will show the way to a uniform blank for the schools and colleges of the country.⁵

A reproduction of the blank, filled in, follows. The use of "C" to show predicted success if a subject is "continued," and of "U" to show ability to "use" it in other fields if it is not continued in college should be noted. "U" is not entered unless the prediction for *continuance* is not high.

THE "JUNIOR YEAR" BLANK

An increasing number of colleges are interested in obtaining information about candidates when they are in the eleventh grade. Information at that time need not be so complete as in the twelfth grade, but it should follow much the same lines.

To supply this need a preliminary report form was also prepared and is in use by the schools.

⁵ An important contribution in this respect has recently been made by the publication of a blank prepared by a committee representing a number of associations. See Appendix, p. 508.

Chapter XIII

STUDY OF THE DEVELOPMENT OF PUPILS IN SUBJECT FIELDS



Departments in the various subject fields studied their objectives more intensively during the early years of the Eight-Year Plan than the teachers concerned, or perhaps any group of teachers, had ever done before. It became evident in this study of objectives that teachers in general, even excellent ones, were not fully aware of any but the most general, and therefore vague, purposes for which they were supposed to be working, and that they often had little appreciation of the importance of the changes that were brought about in their pupils by the experiences of school and out-of-school life. As a matter of fact many an instructor is teaching in his particular subject field (or is teaching at all) only because he found that subject easy and so made a good record in it himself. He assigns a lesson or presents material to his classes, expecting a certain success in learning, but he never looks deeply into his pupils' emotional responses and thought processes or analyzes the developmental stages through which they pass, and the reasons for them.

Because of increased realization of the need for a more analytical approach to the problems of teaching, a demand arose for help in making and keeping teachers aware of the aims for which they should strive. A committee¹ was therefore appointed to investigate methods of recording that might serve such a purpose.

¹ The members of this committee were: Helen M. Atkinson, Genevieve L. Coy, Harry Herron, G. H. B. Melone, Edith M. Penney, Eugene R. Smith, Chairman, Arthur Traxler, John W. M. Rothney. They were assisted by a very large number of school and college teachers who contributed greatly to the undertaking.

The original committee included specialists in various fields, as well as executives. Its first conclusion, resulting from a comparison of objectives of large numbers of teachers, was that, while it did not seem possible to make one form that would be suitable for use in all the fields of knowledge and activity, it would be possible to develop separate forms for those fields that would not only be consistent, but would parallel each other in many respects.

Further experimentation convinced the group that the work should be done largely by specialists in the various fields, assisted by some members of the general group who had studied recording intensively.

The first detailed attack on the problem was made by dividing the original committee, according to its subject interests, into those who would work in English, social studies, mathematics, and science, and by inviting other school and college representatives to join these groups. Meetings usually started with a discussion of the questions involved in the general problem, after which the four groups met separately, coming together again to report progress at the end of the second day.

A very significant development was the increase in breadth of thinking that came to all of the groups, the growth in recognition of the similarity of purposes in different fields, and an appreciation of the importance of common and correlated effort to achieve such purposes. Not only did the groups in mathematics and science spend much time working together, but the mathematics group asked the teachers of social studies to consider a question with them, or some other combination attacked a problem together. After preliminary forms were made, other teachers and schools were asked to criticize them, and eventually through really grueling work carried on with considerable sacrifice by some of the workers, four forms were arrived at.

When this stage was reached, others were invited to join

the committee and forms were added for foreign languages, art, music, physical education, and homemaking.

It was expected that two forms might be needed for foreign languages, one for the modern and the other for the classical languages, but as the work went on it seemed likely that one form could well cover the objectives for both divisions.

Two comments have special significance regarding all the forms. The first is that it proved impossible in any field to limit the objectives to a number that teachers in general would be able to use. The main headings under which judgments can be made are reasonably few, but the sub-heads considered important by the committees increase the possible number of judgments to a point where few teachers would have the time to make so complete a study of their pupils. This may be a strength instead of a weakness, for it brings in enough flexibility to enable any school or teacher to choose the objectives that fit the aims of the institution or the teacher, and to concentrate on the study of their degree of attainment. The record is, then, just as simple, or as extended, as one chooses to make it. It depends absolutely on one's judgment as to which objectives are important enough to justify careful study of each pupil's development in respect to them.

The second comment concerns the "Behavior Description" section on the back of each card. Each committee that analyzed and stated the aims of its department included development in respect to most of the characteristics in the "Behavior Description" list. Each group eventually realized that these characteristics had already been exhaustively studied by a very competent committee, and that there would be no advantage in duplicating that work, even if it were possible to do so. Accordingly, the committees made places for the "Behavior Description" in abbreviated form on their progress cards. It must be understood, however, that this part of the cards can be applied with full effect only through use

of the definitions of characteristics and classifications explained in the Behavior Description section of this report.

A valuable feature of most of the cards is their inclusion of a prediction of future success in the field in question. This is meant to be a basis for the prediction on the "Confidential Report to the Committee on Admissions." Information under "Significant Interests" and the headings following that one are also valuable for transfer as well as for guidance.

The committees endeavored to make these cards as nearly self-explanatory as possible, both in the listing of objectives and the explanation of methods of recording. Here too, however, it must be emphasized that in recording the pupil as high, modal, or low in regard to any objective, the teacher is indicating the kind of growth the pupil is making rather than giving him a mark. The pattern of judgments about the objectives considered should show where the pupil is developing well, and where poorly, and should thus provide data for helping him.

Unfortunately the committees were unable to prepare such cards for all the purposes that might have proved useful. It is likely that the most important omission concerns "core" courses that either include two or more fields, such as English and social studies, or are concerned primarily with the life needs of the pupils. It seems possible, however, that objectives not much different from those that would have been chosen for such a course can be found on the card for "Social Studies," and that this card can therefore be used without serious disadvantage. There have been requests for cards for drama and for instrumental music also, and such cards may yet be devised.

Perhaps in no kind of recording is a teacher likely to be so critical as in that to be used in his own subject, and the less one has studied the detailed objectives in a field, the more likely he is to overlook the implications in such lists as are on these forms. The committees, though they make no

IN

CHOOSING OR USUAL FOR AGE (M), OR LOW (L)
BY CHEAT AT LEAST A FAIRLY DEFINITE OPINION,
MAIN TO INDICATE A SERIOUS LACK.

OBJECT	NOTES
WORK F	
PERSISTE	
EFFECTIV	
SKILL IN	
TECHNIC	
LIBRARY	
ABILITY T	
ABILITY T	
ABILITY T	
ABILITY T	
COMMUN	
COMMUN	
CLEAR	
USE OF V	
AURAL C	
MECHAN	
MECHAN	
MASTERY	
RECOGNIZ	
MAKES AN	
MAKES GEN	
REACHES	
CREATIV	
DRAWN O	
AMOUNT	
CREATIVE	
INDICATE V	
APPREC	
DEVELOP	
DEVELOP	
SENSITIV	
INSIGHT I	
FINDS CL	
SEES IN	
DEVELO	
DEVELOPM	

	YEAR			GRADE		YEAR		
	CREDIT ABLY	BARELY PASSING	FAILING	WITH DISTING TION	VERY WELL	CREDIT ABLY	BARELY PASSING	FAILING
MASTERY								
PREDICT								
READING								
BOOKS OF F	10	11	12	COMMENT				
TYPE OF F								
MEDIAN L								
BOOKS OF								
TYPE OF N								

extravagant claims for their product, hope that *anyone* interested in such forms will take time for careful consideration before deciding that the cards do not quite adequately serve the purposes for which they were designed. It should be noted, for example, that "conscientiousness," which most teachers would expect to find in the list, is not on the front of the card because it is included under "Responsibility-Dependability" in the Behavior Description on the back of the card. Some headings that at first thought seem essential appear in less general form, or are included in more general statements. On the English card, for example, "Skill in obtaining information other than from books," is included, while the more common and *important* (in this field) purpose of obtaining information from books is *omitted*. It is omitted because it is too important and so must appear in more analyzed form. It will be found in such headings as those under "Techniques and Skills" in "Use of Various Reading Techniques," and in the "Reading Record." It is of course included in "Mastery of Essentials of the Course."

To show the method and organization used for these cards the front of the English card is reproduced here.

The back of the card includes, as has been said, the Behavior Description (Chapter X) but uses only the key words, the definitions being omitted. It also has spaces for recording the results of comparable tests, and for making notes about:

Significant Interests, Activities, and Accomplishments

Special Abilities

Significant Limitations

General Comment

The cards in the other subjects follow the same general plan as the English card, but they differ in details in accordance with the particular purposes of the various courses. These differences are not listed because that would require

what would approximate a reproduction of all the cards, and in a rather confusing arrangement. It seems much better for one interested in a particular field to obtain a sample card for that field, in order to study it as a whole.

These cards differ from the ones described in other chapters because while the others are primarily office forms, these are just as definitely teachers' forms, planned to help the teachers in their study of their pupils, and to serve as source material for the other records. From them can be taken the teachers' judgments for entering on the "Behavior Description," and much that goes on the "Form for Transfer from School to College." They serve as a basis for the teachers' reports that become reports to the home. If a cumulative record form is kept, much of the information on it must come from the teachers' cards. It seems, therefore, that these cards, except when data is being taken from them, might well remain in the hands of the teachers, serving as reminders of objectives and offering the opportunity to record information whenever it seems timely.

APPENDICES



Appendix I

CUMULATIVE RECORD FORM

(Prepared by a Committee of the
American Council on Education¹)

As was said in Chapter IX, no work was done by the Committee on Evaluation and Recording on a cumulative record form for the use of school offices because the American Council on Education was planning to revise the form that had been used so widely since its publication in 1930. The revision for secondary schools has now been completed and the card can be obtained from the Council's office in Washington. It accords with the principles and methods of the other forms described in this volume, and so fits well into the set from which a school can choose its equipment for recording.

The cumulative record form is a double sheet of tagboard that fits an 8½" by 11" file. It furnishes space for all the commonly recorded facts about a pupil and his family, and for a six-year history of his school career.

One of the largest spaces on the card is given to the history and analysis of the pupil's progress in subject fields. This allows opportunity for whatever type of reporting a school uses, though the directions suggest some form of analysis such as is described in Chapter XI. Alternative forms provide for recording test results in tabular or graphic form, and there is also provision for interpreting the test record in relation to the pupil's academic achievement.

The "Description of Behavior" section uses material from the card and manual described in Chapter X, and adds spaces for advice by guidance officers, and for follow-up after the pupil leaves school.

¹ Richard D. Allen, Associate Superintendent, Providence, R. I.; Millard E. Gladfelter, Temple University; William S. Learned, Carnegie Foundation

UNIFORM COLLEGE ENTRANCE BLANK

In 1941 under the joint auspices of The American Council on Education and The National Association of Secondary School Principals a committee was appointed representing these associations, and the New England Association of Colleges and Secondary Schools; the Middle States Association of Colleges and Secondary Schools; the North Central Association of Colleges and Secondary Schools; the Southern Association of Colleges and Secondary Schools; the Progressive Education Association; the American Association of Collegiate Registrars, for the purpose of considering the demand for an improved and uniform college entrance blank. The chairman and secretary of the Committee on Evaluation and Recording were members.

This committee considered blanks already prepared by various groups and agreed upon a form which has now been published by the National Association of Secondary School Principals and can be obtained from its office in Washington, D. C.

While this form is much more condensed than that prepared for the Eight-Year Study, having in particular a limited space for free comment about the candidate, it has much in common with that form and recognizes much the same educational principles. It offers opportunity for the use of analyses or predictions instead of marks if a school prefers them, omits any reference to units, provides space for annual tests, and gives emphasis to the description of behavior.

This form shows marked progress toward present-day objectives and promises to influence school and college relations constructively.

for the Advancement of Teaching; John W. M. Rothney, Wisconsin University, Secretary; Donald J. Shank, Assistant to the President, American Council on Education; Eugene R. Smith, The Beaver Country Day School, Chairman; Arthur E. Traxler, Educational Records Bureau; Edmund G. Williamson, University of Minnesota; Ben Wood, Cooperative Test Service.

Appendix II

TABLES FOR CHAPTER II

TABLE 1

Means and Standard Deviations of Per Cent Scores, Reliability Coefficients for each Category of Form 2.52 for Sample School Populations for Each of Grades Nine, Ten, Eleven and Twelve

GRADE 9

	School A			School E			School F			School G		
	N	M	σ	N	M	σ	N	M	σ	N	M	σ
General accuracy	52	34.9	10.6	91	100	42.5	9.2	86	15	48.0	9.8	.96
Beyond data	52	54.9	9.8	86	100	48.2	13.0	.92	15	31.8	15.4	.99
Caution	52	27.8	8.8	78	100	18.0	11.0	.90	15	36.8	12.7	.96
Crude errors	52	19.3	9.4	86	100	8.8	3.9	.48	15	10.9	7.0	.87
True false	52	43.4	22.7	89	100	64.0	8.9	.54	15	76.2	13.9	.79
Insufficient data	52	30.8	13.8	85	100	44.2	16.6	.90	15	62.6	17.7	.97
Probably true—probably false	52	25.8	8.5	62	100	36.0	16.7	.88	15	19.4	9.2	.80
											20.8	7.4
												.59

TABLE 1—(Continued)

GRADE 10

	School A		School B		School C		School F		School G							
General accuracy	44	41.4 10.2	90	33	47.8 10.0	.91	35	50.1	7.3	.83	25	54.2 12.5	.98	24	48.0 10.0	.93
Beyond data	44	51.8 13.0	.93	33	47.4 10.5	.91	35	46.9	8.6	.86	25	20.0 8.2	.89	24	44.8 11.6	.95
Caution	44	25.4 11.0	.89	33	22.2 11.2	.91	35	23.3	8.8	.85	25	35.6 10.5	.91	24	24.9 8.3	.86
Crude errors	44	15.8 5.6	.65	33	14.1 5.5	.69	35	10.8	4.9	.72	25	7.7 3.6	.47	24	11.5 4.4	.62
True—false	44	55.2 14.1	.74	33	60.3 13.2	.73	35	68.5 10.4	.67	.67	25	79.6 12.0	.70	24	54.9 14.1	.81
Insufficient data	44	37.8 15.7	.89	33	42.8 14.1	.88	35	48.3 12.3	.86	.86	25	66.5 21.9	.99	24	44.5 14.8	.91
Probably true—probably false	44	29.7 13.4	.86	33	42.5 17.1	.91	35	33.8 14.0	.86	.86	25	19.4 11.5	.88	24	35.1 12.5	.83

GRADE 11

	School A		School B		School E		School F		School G									
General accuracy	50	47.6 9.0	.88	29	52.0 10.4	.92	79	60.3	8.3	.86	32	55.8	8.0	.87	33	49.9	7.1	.80
Beyond data	50	43.8 12.2	.94	29	42.9 13.0	.95	79	40.8	9.8	.89	32	18.9	7.4	.86	33	45.4	10.0	.90
Caution	50	30.5 12.5	.91	29	24.2 11.5	.91	79	15.7	8.4	.86	32	37.4	12.4	.91	33	24.0	11.4	.92
Crude errors	50	13.7 5.1	.65	29	12.7 3.9	.47	79	10.1	3.1	.28	32	8.9	5.2	.72	33	11.4	4.9	.70
True—false	50	58.9 12.1	.71	29	61.2 11.2	.68	79	70.8	8.8	.61	32	73.1	16.7	.83	33	69.7	8.3	.56
Insufficient data	50	49.0 12.5	.84	29	51.1 18.6	.90	79	53.6	11.0	.82	32	75.4	11.5	.86	33	47.4	12.6	.85
Probably true—probably false	50	30.9 14.9	.88	29	39.1 15.4	.88	79	54.1	14.4	.83	32	27.8	11.5	.84	33	34.4	14.1	.86

GRADE 12

	<i>School A</i>		<i>School B</i>		<i>School D</i>		<i>School F</i>		<i>School G</i>											
General accuracy	45	49.7	7.3	.82	35	54.5	10.4	.92	97	49.6	9.6	.88	27	63.6	7.8	.89	30	56.6	10.1	.92
Beyond data	45	44.8	8.6	.86	35	40.6	12.1	.95	97	33.2	10.7	.89	27	13.3	9.0	.94	30	39.4	15.9	.99
Caution	45	24.3	8.8	.84	35	23.7	10.2	.92	97	28.2	13.9	.92	27	37.0	16.9	.98	30	23.3	11.8	.93
Crude errors	45	12.8	4.9	.65	35	11.9	3.8	.47	97	10.9	4.6	.54	27	4.7	3.7	.69	30	9.6	4.0	.64
True—false	45	63.7	10.0	.59	35	61.6	10.5	.62	97	73.0	16.5	.80	27	85.5	11.9	.76	30	69.2	11.4	.77
Insufficient data	45	48.9	11.2	.82	35	62.6	11.5	.83	97	61.5	14.7	.88	27	87.4	11.2	.91	30	51.4	13.9	.89
Probably true—probably false	45	34.7	14.0	.85	35	41.6	16.3	.89	97	24.6	16.3	.89	27	28.2	18.4	.94	30	47.0	16.6	.90

TABLE 2

Correlations between Scores on Form 2.51 (Corrected for Attenuation) for 284 Pupils in Two Large Public High Schools not in the Eight-Year Study

Score	Accuracy with True-False	Accuracy with Probably True and Probably False	Accuracy with Insufficient Data	Beyond Data	Caution	Crude Errors
General Accuracy	.766	.650	.786	-.734	-.132	-.880
Accuracy with true-false		.470	.314	.075	-.359	-.882
Accuracy with probably true and probably false			-.002	-.052	-.741	-.301
Accuracy with insufficient data				-.981	.491	-.719
Beyond data					-.590	.637
Caution						-.150

TABLE 3

Means, Standard Deviations, Reliability Coefficients for Each Category of Form 2.71 for Five School Populations for each of Grades Seven, Eight and Nine

GRADE 7

Score	School A				School B				School F				School G			
	M		σ		M		σ		M		σ		M		σ	
	N	r	N	r	N	r	N	r	N	r	N	r	N	r	N	r
General accuracy	68	48.5	10.5	.80	26	53.7	13.2	.92	22	66.5	15.2	.96	29	54.7	14.0	.92
Beyond data	68	54.4	17.7	.86	26	53.8	16.9	.88	22	32.5	22.2	.96	29	43.4	26.0	.97
Caution	68	27.7	12.6	.78	26	18.3	10.4	.78	22	14.5	7.9	.70	29	25.1	16.2	.90
Crude error	68	19.7	6.8	.42	26	20.0	9.4	.70	22	18.8	7.8	.58	29	9.3	8.8	.65
Agree—disagree	26	61.0	14.7	.85	22	67.2	12.3	.78	29	54.3	15.2	.84
Uncertain	26	46.6	16.9	.89	22	67.0	22.5	.97	29	55.7	25.1	.96

GRADE 8

Score	School A				School B				School D				School F			
	M		σ		M		σ		M		σ		M		σ	
	N	r	N	r	N	r	N	r	N	r	N	r	N	r	N	r
General accuracy	60	50.6	12.3	.87	27	65.0	11.2	.89	24	54.4	12.5	.91	23	55.8	11.3	.88
Beyond data	60	63.2	18.9	.88	27	38.2	17.2	.90	24	52.2	17.3	.90	23	41.1	16.9	.90
Caution	60	14.7	9.8	.73	27	15.9	7.5	.63	24	19.8	12.6	.86	23	19.2	11.9	.84
Crude error	60	19.7	7.9	.56	27	14.6	6.2	.48	24	17.8	6.3	.40	23	15.3	8.0	.68
Agree—disagree	27	68.7	9.0	.61	24	62.2	15.4	.87	23	57.6	15.2	.87
Uncertain	27	63.4	14.7	.85	24	48.0	17.9	.91	23	53.0	14.3	.84

GRADE 9

Score	School A				School C				School E				School F			
	M		σ		M		σ		M		σ		M		σ	
	N	r	N	r	N	r	N	r	N	r	N	r	N	r	N	r
General accuracy	65	59.7	15.3	.91	28	59.9	16.1	.96	23	64.0	13.9	.97	34	68.4	11.5	.90
Beyond data	65	46.5	24.0	.92	28	32.7	16.2	.90	23	23.6	10.4	.81	34	29.1	14.8	.88
Caution	65	15.6	10.2	.80	28	11.8	8.5	.78	23	14.6	8.2	.74	34	14.1	9.5	.79
Crude error	65	17.0	7.3	.55	28	14.3	8.2	.68	23	13.6	5.0	.43	34	16.6	8.8	.70
Agree—disagree	28	72.7	13.0	.83	23	70.0	11.7	.78	34	67.9	13.3	.82
Uncertain	28	66.0	15.8	.89	23	76.2	10.3	.81	34	68.9	13.5	.84

TABLE 4

*Correlations between Certain Scores on Form 1.3b for 283 Pupils
in Two Schools in the Eight-Year Study*

Score		Accurate Conclusions	Right Reasons	Percent Reasons Right	Number Right Principles	Number Right Controls	Number Analogies	Number Right Analogies	Percent Ridicule, Tel., A. C.
	Column	1	8	9	12	16	18	19	25
Uncertain rea- sons, lack of knowledge	5	.34							
Number reasons	7	.15							
Right reasons	8	.53							
Percent reasons right	9	.68							
Number principles	11	.42							
Number right principles	12	.53	.85						
Percent principles right	13			.62					
Number right controls	16		.38		.04				
Number analogies	18	.17							
Number right analogies	19	.21	.70		.59	.01			
Number authorities	21						.61		
Number right authorities	22		.58		.40	.01		.54	
Percent ridicule, Tel., A. C.	25			-.46					
Percent inconsistent	27			-.20					.24

TABLE 5
Reliability Coefficients, Means, and Standard Deviations for Application of Principles of Science, Form 1.3b

Statistic	Form 1.3b N = 283		Form 1.3b N = 100		Form 1.3c N = 100		r*
	Reliability	Mean	σ	Mean	σ	Mean	
Accurate conclusions	.36	4.3	1.68	4.0	1.6	4.1	.48
Too uncertain conclusions	.32	1.7	1.30				
Too certain conclusions	.09	1.1	0.73				
General accuracy	.83	27.8	19.60				
Uncertain reasons, lack of knowledge	.62	3.01	2.48				
Number reasons	.87	31.2	13.10	31.4	13.60	28.9	.79
Right reasons	.90	18.0	8.08	16.6	8.18	16.3	.67
Percent reasons right	.75	58.5	16.90	53.7	17.10	58.8	.56
Number right principles	.77	14.0	5.98				
Number right principles		11.3	5.38				
Percent principles right		77.2	15.66	73.6	19.68	79.9	.36
Number controls	.81	6.6	5.60				
Number right controls	.75	2.6	2.90				
Number analogies	.61	3.7	2.64				
Number right analogies	.48	2.3	1.71				
Number authorities	.66	2.7	2.44				
Number right authorities	.51	1.9	1.64				
Number ridicule, Tel., A. C.		4.4	3.26				
Percent ridicule, Tel., A. C.		13.1	7.38	12.1	6.62	12.9	.33
Number inconsistent		2.5	2.84				
Percent inconsistent		7.9	7.41	7.6	7.92	7.1	.15

* Product-moment correlation coefficient between Form 1.3b and Form 1.3c.

TABLE 6
*Reliabilities, Intercorrelations, Means and Standard Deviations for Form 5.12 Based on
 the Scores of 351 Pupils in Public High Schools**

	Column	1	2	3	4	5	6	7	8	9	10
Right conclusions	1	.23	-.96								
Wrong conclusions	2		.28								
Relevant as relevant	3			.72							
Irrelevant as relevant	4				-.23	.30	-.64				
Irrelevant as irrelevant	5				.83	-.80	.15				
Relevant as irrelevant	6					.88	.73				
Accuracy with definitions	7							.81			.25
Indirect argument	8								.38	.33	.43
Ad hominem	9								.80	.48	.27
If—then	10									.73	.72
Mean		4.5	3.5	54.2	27.3	62.3	35.0	13.7	17.1	15.8	14.2
Standard deviation		1.55	1.55	15.6	13.8	18.2	15.3	5.9	5.4	4.7	4.8

* Reliabilities are given in the main diagonal of the table.

TABLE 7

*Reliabilities, Intercorrelations, Means and Standard Deviations for Form 5.22 Based on the Scores of 307 Pupils in Public High Schools**

Score	Column	1	2	3	4	5	6	7	8	9	10	11	12	13
% Called Supporting which were	1	.80												
	2	.72	.61		.59			.59	.01		.48	.50	.44	.19
	3	.80			.43	.63								
% Called Contradictory which were	4				.82			.47	.10		.27	.36	.32	.05
	5					.77	.56							
	6						.79							
% of Supporting called Critical which were	7							.74			.36	.04	.36	.08
	8								.52					
Number of Conclusions which were	9									.20				
% called Relevant which were	10										.82	.74	.74	.29
	11													
% of Relevant called Practical which were	12												.71	.51
	13													
Mean		73	23	16	66	33	15	37	18	4.8	71	45	55	27
Standard Deviation		13	13	15	19	17	10	16	12	1.5	14	17	17	13

* Reliabilities are given in the main diagonal of the table.

Appendix III

TABLES FOR CHAPTER III

TABLE 1

Means, Standard Deviations, and Reliabilities for Test 1.41

	Grade 10			Grade 11			Grade 12			Total		
	Mean	Sigma	r	Mean	Sigma	r	Mean	Sigma	r	Mean	Sigma	r
Total Reasons ¹	54.2	11.1	.82	55.3	13.7	.87	46.8	11.9	.89	51.8	12.9	.87
Accurate Reasons ¹	37.5	8.3	.85	39.8	9.5	.89	36.5	9.0	.87	37.9	9.0	.87
Ratio ¹	4.5	.98	.82	4.9	1.1	.87	4.5	1.1	.84	4.6	1.1	.85
No. Inconsistent ¹	6.4	4.2	.78	5.7	4.3	.76	3.5	2.6	.65	5.1	3.9	.77
% Inconsistent ¹	9.1	8.0	.80	7.5	7.5	.48	5.5	5.5	.72	7.2	7.2	.50
Untenable ²	6.2	2.5	.35	6.2	2.6	.44	4.4	2.5	.52	5.5	2.7	.50
Irrelevant ²	3.9	1.9	.35	2.2	2.2	.27	1.7	1.7	.33	3.3	2.0	.20
Undemocratic Values ² ...	6.4	4.7	.84	5.5	4.8	.86	3.8	4.2	.86	5.2	4.7	.86
Democratic Values ²	22.3	8.6	.90	25.4	8.9	.91	23.7	9.3	.92	23.8	9.0	.91
Rationalization ²	8.2	3.1	.54	7.7	3.7	.72	5.9	3.0	.63	7.2	3.4	.67
% Democratic Values ²										62.9	16.4	.70

¹ Computed by split-half method.

² Computed by Kuder-Richardson formula.

³ Computed by correlating two forms of the test 1.41 and 1.42.

TABLE 2
Reliability Coefficients for Test 4.21-4.31

	9th Grade (108)	10th Grade (145)	11th Grade (169)	12th Grade (179)	Total (601)
<i>Liberalism</i>					
D Col. 1.....	.74	.78	.78	.80	.79
ER 2.....	.77	.78	.80	.84	.81
LU 3.....	.80	.83	.85	.86	.84
R 4.....	.81	.84	.88	.86	.86
N 5.....	.66	.75	.79	.80	.77
M 6.....	.79	.84	.88	.86	.86
<i>Conservatism</i>					
D Col. 7.....	.62	.70	.76	.72	.74
ER 8.....	.71	.73	.81	.78	.78
LU 9.....	.70	.77	.79	.78	.77
R 10.....	.83	.77	.83	.77	.81
N 11.....	.66	.72	.75	.69	.72
M 12.....	.71	.79	.82	.80	.80
<i>Uncertainty</i>					
D Col. 13.....	.72	.86	.86	.83	.85
ER 14.....	.81	.82	.82	.82	.82
LU 15.....	.82	.84	.85	.83	.84
R 16.....	.77	.74	.79	.83	.79
N 17.....	.78	.82	.80	.81	.81
M 18.....	.84	.84	.85	.83	.84
<i>Consistency</i>					
D Col. 19.....	.54	.42	.57	.32	.56
ER 20.....	.42	.42	.56	.51	.51
LU 21.....	.48	.57	.57	.57	.57
R 22.....	.44	.44	.58	.58	.54
N 23.....	.23	.38	.51	.54	.46
M 24.....	.32	.55	.68	.65	.61
<i>Totals</i>					
Liberalism.....	.93	.94	.95	.95	.95
Conservatism.....	.91	.92	.94	.92	.93
Uncertainty.....	.96	.96	.96	.96	.96
Consistency.....	.82	.78	.87	.88	.85

TABLE 3
Means and Standard Deviations for Test 4.21-4.31

	Grade 9			Grade 10			Grade 11			Grade 12			Total							
	M		σ	M		σ	M		σ	M		σ	M		σ					
	raw	%		raw	%		raw	%		raw	%		raw	%						
			raw			%			raw			%			raw	%	raw	%	raw	%
<i>Liberalism</i>																				
D.....	59.3	29.7	12.5	6.3	59.7	29.9	13.6	6.8	57.3	28.7	14.1	7.1	65.2	32.6	14.0	7.0	60.6	30.3	14.1	7.0
ER.....	50.2	15.1	17.4	5.2	45.4	13.6	17.9	5.4	45.5	13.7	18.7	5.6	56.2	16.9	21.1	6.3	49.5	14.85	19.6	5.9
LU.....	56.0	20.2	17.2	6.2	56.7	20.4	18.8	6.8	54.2	19.5	20.3	7.3	66.3	23.9	19.8	7.1	58.7	21.1	19.3	7.0
R.....	68.8	16.5	20.9	5.0	63.2	15.2	23.7	5.7	57.5	13.8	26.8	6.4	68.1	16.3	25.3	6.1	64.1	15.4	25.1	6.0
N.....	48.9	17.6	12.9	4.6	50.7	18.3	15.0	5.4	50.0	18.0	16.7	6.0	56.5	20.4	16.1	5.8	51.9	18.7	15.8	5.5
M.....	54.7	21.9	16.4	6.6	56.1	22.4	18.8	7.5	53.2	21.3	21.4	8.6	65.5	26.2	19.6	7.8	57.6	23.0	20.3	8.1
<i>Conservatism</i>																				
D.....	22.9	11.5	8.9	4.5	20.8	10.4	9.7	4.9	19.8	9.9	11.3	5.7	16.8	8.4	9.5	4.8	19.8	9.9	10.7	5.2
ER.....	26.8	8.0	14.1	4.2	27.1	8.1	14.7	4.4	24.5	7.3	17.1	5.1	22.1	6.6	15.3	4.6	25.0	7.5	16.1	4.8
LU.....	22.1	8.0	12.1	4.4	20.6	7.4	13.7	4.9	18.9	6.8	14.1	5.1	15.8	5.7	12.5	4.5	19.0	6.8	13.4	4.8
R.....	15.7	3.8	18.0	4.3	17.5	4.3	17.1	4.1	21.3	5.1	19.6	4.7	14.2	3.4	15.4	3.7	17.4	4.2	17.7	4.3
N.....	31.0	11.2	11.9	4.3	25.2	9.1	12.8	4.6	25.3	9.1	13.6	4.9	23.3	8.4	11.4	4.1	25.7	9.3	12.8	4.6
M.....	25.7	10.3	12.0	4.8	22.8	9.1	13.9	5.6	23.1	9.2	15.8	6.3	16.72	6.7	13.3	5.3	21.5	8.6	14.4	5.7

<i>Uncertainty</i>																				
D.....	18.0	9.0	12.0	6.0	19.7	9.9	14.4	7.2	23.0	11.5	16.0	8.0	17.7	8.9	12.4	6.2	19.8	9.9	14.0	7.0
ER.....	23.8	7.1	17.4	5.2	27.8	8.3	18.8	5.6	30.2	9.1	19.2	5.8	22.3	6.7	17.3	5.2	26.1	7.8	18.5	5.5
LU.....	21.8	7.8	16.0	5.8	22.9	8.2	17.0	6.1	26.2	9.4	18.8	6.8	18.1	6.5	14.9	5.4	22.2	8.0	17.1	6.2
R.....	15.0	3.6	15.1	3.6	19.8	4.8	15.0	3.6	22.1	5.3	17.8	4.3	17.6	4.2	18.1	4.3	18.7	4.5	16.8	4.0
N.....	21.0	7.6	14.0	5.0	23.1	8.3	15.7	5.7	25.2	9.1	16.0	5.8	20.0	7.2	15.0	5.4	22.4	8.1	15.4	5.6
M.....	21.7	8.7	15.7	6.3	22.8	9.1	16.4	6.6	24.9	10.0	17.6	7.0	18.3	7.3	14.1	5.6	19.8	7.9	16.2	5.5
<i>Consistency</i>																				
D.....	52.7	13.2	14.3	3.6	58.5	14.6	13.0	3.3	55.9	14.0	15.1	3.8	61.0	15.3	14.5	3.6	57.5	14.4	14.6	3.6
ER.....	48.9	7.3	16.5	2.5	50.7	7.6	16.4	2.5	50.6	7.6	19.0	2.9	56.7	8.5	17.9	2.7	50.2	7.5	17.9	2.7
LU.....	52.3	9.4	15.6	2.8	53.6	9.6	17.5	3.2	55.2	9.9	17.6	3.2	62.4	11.2	17.0	3.1	56.4	10.2	17.5	3.2
R.....	66.4	8.0	18.4	2.2	64.0	7.7	18.5	2.2	61.2	7.3	21.7	2.6	68.8	8.3	20.8	2.5	65.0	7.8	20.3	2.4
N.....	45.6	8.2	12.9	2.3	50.4	9.1	14.5	2.6	49.7	8.9	16.0	2.9	51.2	9.2	16.7	3.0	50.5	9.1	15.5	2.8
M.....	54.4	10.9	13.4	2.7	58.4	11.7	16.1	3.2	57.6	11.5	18.9	3.8	64.6	12.9	18.0	3.6	60.2	12.0	17.5	3.5
<i>Totals</i>																				
Liberalism.....	55.1	11.0	12.1	24.2	54.6	109.2	12.7	25.4	52.6	105.2	15.0	30.0	62.1	124.2	14.8	29.6	56.4	112.9	14.5	28.9
Conservatism.....	24.1	48.2	9.5	19.0	23.0	46.0	9.9	19.8	22.2	44.4	11.8	23.6	18.6	37.2	9.3	18.6	21.7	43.4	10.4	20.8
Uncertainty.....	20.8	41.6	13.1	26.2	22.4	44.8	14.4	28.8	25.1	50.2	15.0	30.0	24.7	49.4	12.6	25.2	22.0	43.9	14.0	28.0
Consistency.....	51.4	51.4	11.3	11.3	55.0	55.0	10.3	10.3	54.7	54.7	13.5	13.5	59.4	59.4	13.5	13.5	55.6	55.6	12.7	12.7

TABLE 4

Intercorrelations of Certain Scores on Scale of Beliefs 4.21-4.31

Score	Liberalism					Conservatism					Totals		
	D	ER	LU	R	N	D	ER	LU	R	N	Lib.	Con.	Unc.
	1	2	3	4	5	7	8	9	10	11	25	26	27
Liberalism													
ER Col. 2	.59												
LU 3	.70	.64											
R 4	.59	.30	.54										
N 5	.64	.43	.52	.54									
M 6	.59	.33	.51	.52	.76								
Conservatism													
ER 8						.57							
LU 9						.71	.62						
R 10						.61	.31	.59					
N 11						.61	.42	.61	.57				
M 12						.52	.32	.48	.45	.73			
Totals													
Lib. 25											— .37		
Con. 26											— .69	— .40	
Unc. 27													
Consi. 28											.66	.42	— .33

Appendix IV

TABLES FOR CHAPTER IV

Students from a large public senior high school are the only ones who have taken the final revised form 3.32. Eleven classes, distributed as follows, constituted the population.

TABLE 1

Grade	Boys	Girls	Total
10	56	59	115
11	46	56	102
12	52	66	118
Total	154	181	335

TABLE 2

Means, Standard Deviations, and Estimates of Reliability of "Appreciation" Scores on Parts I, II, and III of Form 3.32

	Mean	σ	r
<i>Part I (35 items)</i>			
Grade 10.....	57.0	17.89	.85
Grade 11.....	61.8	17.09	.84
Grade 12.....	66.6	17.66	.85
<i>Part II (40 items)</i>			
Grade 10.....	47.0	18.18	.86
Grade 11.....	52.4	18.84	.88
Grade 12.....	55.4	17.73	.86
<i>Part III (25 items)</i>			
Grade 10.....	49.7	17.09	.73
Grade 11.....	57.0	19.32	.80
Grade 12.....	53.6	17.38	.77
<i>Total (100 items)</i>			
Grade 10.....	61.6	15.75	.92
Grade 11.....	51.6	16.84	.94
Grade 12.....	53.3	14.33	.91

TABLE 3

*Means, Standard Deviations, and Estimates of Reliability of "Non-Appreciation"
Scores on Parts I, II, and III of Form 3.32*

	Mean	σ	r
<i>Part I</i> (35 items)			
Grade 10.....	36.8	17.28	.84
Grade 11.....	32.3	16.26	.83
Grade 12.....	29.2	17.01	.85
<i>Part II</i> (40 items)			
Grade 10.....	47.8	17.08	.84
Grade 11.....	40.8	18.60	.88
Grade 12.....	39.5	15.45	.82
<i>Part III</i> (25 items)			
Grade 10.....	41.5	17.90	.77
Grade 11.....	34.2	17.76	.79
Grade 12.....	39.2	15.94	.74
<i>Total</i> (100 items)			
Grade 10.....	42.3	15.24	.92
Grade 11.....	36.1	16.35	.94
Grade 12.....	35.0	14.53	.92

TABLE 4

Means, Standard Deviations, and Estimates of Reliability of "Uncertain" Scores on Parts I, II, and III of Form 3.32

	Mean	σ	r
<i>Part I (35 items)</i>			
Grade 10.....	8.0	8.98	.81
Grade 11.....	7.9	7.63	.78
Grade 12.....	6.1	6.75	.74
<i>Part II (40 items)</i>			
Grade 10.....	7.5	8.01	.79
Grade 11.....	8.8	9.21	.84
Grade 12.....	7.5	8.39	.80
<i>Part III (25 items)</i>			
Grade 10.....	8.4	11.67	.79
Grade 11.....	10.3	11.89	.83
Grade 12.....	10.1	11.82	.77
<i>Total (100 items)</i>			
Grade 10.....	8.2	8.45	.92
Grade 11.....	8.5	8.50	.93
Grade 12.....	7.8	8.17	.92

TABLE 5

Means, Standard Deviations, and Estimates of Reliability of "Appreciation" Scores on Parts IIA, IIB, IIC, and IID of Form 3.32

	Mean	σ	r
<i>Part IIA</i> (10 items)			
Grade 10.....	67.2	20.01	.51
Grade 11.....	73.6	21.14	.65
Grade 12.....	73.8	21.60	.66
<i>Part IIB</i> (10 items)			
Grade 10.....	32.1	23.57	.73
Grade 11.....	36.5	25.10	.75
Grade 12.....	41.4	25.67	.74
<i>Part IIC</i> (10 items)			
Grade 10.....	48.8	24.01	.68
Grade 11.....	52.2	24.99	.73
Grade 12.....	54.1	20.21	.59
<i>Part IID</i> (10 items)			
Grade 10.....	54.8	23.85	.67
Grade 11.....	60.9	25.45	.73
Grade 12.....	65.8	25.66	.75

TABLE 6

Means, Standard Deviations, and Estimates of Reliability of "Non-Appreciation" Scores on Parts IIA, IIB, IIC, and IID of Form 3.32

	Mean	σ	r
<i>Part IIA (10 items)</i>			
Grade 10.....	30.7	19.23	.53
Grade 11.....	23.9	16.69	.49
Grade 12.....	8.4	18.00	.56
<i>Part IIB (10 items)</i>			
Grade 10.....	68.4	24.77	.72
Grade 11.....	61.3	28.32	.80
Grade 12.....	58.7	27.98	.79
<i>Part IIC (10 items)</i>			
Grade 10.....	56.5	24.40	.69
Grade 11.....	51.7	24.55	.72
Grade 12.....	50.4	20.07	.58
<i>Part IID (10 items)</i>			
Grade 10.....	48.5	24.19	.69
Grade 11.....	39.6	23.04	.68
Grade 12.....	36.8	23.64	.73

TABLE 7

LIST OF PAINTINGS USED IN THE TEST

No.	Artist	Name of the Painting	Collection—Catalogue
1.	Picasso	The Absinth-drinker	Hamburg Museum
2.	Michelangelo	Head of Adam	(Detail) Creation of Adam—Sistina, Rome
3.	Cézanna	Peasant	Conger Goodyear, New York, (Venturi, No. 687)
4.	Corot	Girl with Pearl	Louvre (Robaut, No. 1507)
5.	van Gogh	Self Portrait	V. W. van Gogh—Amsterdam (De la Faille No. 344)
6.	Vermeer	Portrait of a Young Girl	Hague, Royal Gallery (Hofstede, No. 44)
7.	van Gogh	Self Portrait	Museum, Amsterdam (De la Faille, No. 522)
8.	Rembrandt	Self Portrait	Kunsthist. Museum, Vienna (Hofstede, No. 580)
9.	Dürer	Self Portrait	Pinakothek-Muenchen (Tietze No. 164)
10.	Mainardi	Portrait of a Young Man	K. Friedrich Museum, Berlin (Cat. No. 86)
11.	Breughel	The Winter	Kunsthist. Museum, Vienna (deLoo A 24)
12.	Rembrandt	A Boy Reading	Kunsthist. Museum, Vienna (Hofstede No. 238)
13.	El Greco	View of Toledo	Metropolitan Museum, New York (A. L. Mayer, No. 315)
14.	Hals	A Fool with a Mandolin	G. de Rothschild, Paris (Hofstede No. 98)
15.	Gauguin	Farm at the Pouldu	Collection Vollard, Paris
16.	Breughel	The Summer	Metropolitan Museum, New York (deLoo A 25)
17.	Corot	Paysage	Louvre, Paris (Robaut, No. 1625)
18.	Kokoschka	Towerbridge, London	Museum, Hamburg
19.	Rembrandt	Jakob blessing Joseph's sons	Gallery, Cassel (Hofstede, No. 22)
20.	Gauguin	Landscape in Brittany	Collection Mesnard, Paris
21.	Dürer	Self Portrait	Prado, Madrid (Tietze, No. 152)
22.	van Gogh	Dr. Gachet	Gallery, Frankfurt M. (De la Faille No. 753)
23.	Lorenzo di Credi	Portrait of a Girl	K. Friedrich Museum, Berlin (Cat. No. 80)

<i>No.</i>	<i>Artist</i>	<i>Name of the Painting</i>	<i>Collection—Catalogue</i>
24.	Picasso	The Guitarist	Art Institute, Chicago (Zervos: Picasso 1895-1906, No. 202)
25.	Cézanne	The Card Players	Louvre, Paris (Venturi No. 558)
26.	Vermeer	The Kitchenmaid	Collection Six, Amsterdam (Hofstede, No. 17)
27.	Dürer	Hieronymus Holzschuher	German Museum, Berlin (Tietze No. 957)
28.	Corot	Interrupted Reading	Art Institute, Chicago (Robaut, No. 1431)
29.	El Greco	St. Martin and the Beggar	Art Institute, Chicago (A. L. Mayer, No. 298)
30.	van Gogh	Pear Tree in Blossoms	Collection V. W. van Gogh, Amsterdam (De la Faille No. 405)
31.	Hals	A Mulatto	Museum, Leipzig (Hofstede No. 96)
32.	Cézanne	A Village	Collection George Renard, Paris (Venturi No. 307)
33.	Breughel	The Autumn	Kunsthist. Museum, Vienna (deLoo A 26)
34.	Kokoschka	Flowers on the Window	Munich
35.	El Greco	Mater Dolorosa	Munich (A. L. Mayer, No. 86)
36.	van Gogh	Blossoming Almond Spray	Collection V. W. van Gogh, Amsterdam (De la Faille No. 392)
37.	Michelangelo	Adam, Creation of Adam	(Detail) Sistina, Rome
38.	Rembrandt	A Young Girl at an Open Half Door	Art Institute, Chicago (Hofstede No. 324)
39.	Cézanne	Basket of Apples	Art Institute, Chicago (Venturi No. 600)
40.	Hals	The Gipsy Girl	Louvre, Paris (Hofstede No. 119)

TABLE 8

LIST OF PAINTINGS USED IN THE COMPARABLE FORM

<i>No.</i>	<i>Artist</i>	<i>Name of the Painting</i>	<i>Collection—Catalogue</i>
1.	Breughel	The Peasants' Wedding	Kunsthist. Museum, Vienna (deLoo A 27)
2.	Bronzino	Bia de Medici	Uffizi, Florenz (A McComb, p. 61)
3.	van Gogh	Sun Flowers	Collection V. W. van Gogh Amsterdam (De la Faille 458)
4.	Rembrandt	Self Portrait	Louvre, Paris (Hofstede No. 569)
5.	Roger v. d. Weyden	The Knight with the Arrow	Museum, Brussels
6.	Ambrogio da Predis	Portrait (Beatrice d'Este)	Ambrosiana, Milan
7.	Modersohn-Becker	Still-life with Flowers	Museum, Hamburg
8.	Breughel	Fight of Lent with Carnival	(Detail) Kunsthist. Museum, Vienna (deLoo A 2)
9.	Gauguin	The Girl with the Fan	Folkwang Museum, Essen
10.	Michelangelo	Head of the Prophet Jeremiah	Sistina, Rome
11.	El Greco	Cardinal Fernando Nino Guevara	Metropolitan Museum, New York (A. L. Mayer, No. 331)
12.	Memling	Portrait Nicolas di Sforzore	Spinelli Museum, Antwerp (Weale: Memling p. 13)
13.	Cézanne	The Smoker	Kunsthalle, Mannheim (Venturi, 684)
14.	Vermeer	A Lady at the Virginals	Royal Collection Windsor (Hofstede, No. 28)
15.	M. Laurencin	Portrait of a Girl	Pallas Gallery
16.	Cézanne	Vase of Tulips	Art Institute, Chicago (Venturi, 617)
17.	R. Dufy	Window in Nice	Art Institute, Chicago
18.	van Gogh	Portrait of an Old Peasant	Collection Bernheim jeune Paris (De la Faille 444)
19.	Carl Hofer	Girls Throwing Flowers	Art Institute, Chicago
20.	van Gogh	The Zouave	Collection Unger = Mens, Rotterdam (De la Faille 424)
21.	Dégas	Woman Drying her Neck	Louvre, Paris
22.	Dégas	Girls Ironing	Louvre, Paris
23.	Modersohn-Becker	Still-life with Fruits	
24.	Breughel	The Unfaithful Shepherd	Pennsylvania Museum of Art (deLoo A 29)

<i>No.</i>	<i>Artist</i>	<i>Name of the Painting</i>	<i>Collection—Catalogue</i>
25.	Goya	The Bandit Margato, Shot	Art Institute, Chicago (A. L. Mayer, No. 597c)
26.	Winslow Homer	The Gulf Stream	Art Institute, Chicago
27.	Rousseau, Henri	The Cascade	Art Institute, Chicago
28.	Cézanne	Man in a Cotton Cap	Museum of Modern Art, New York (Venturi, 73)
29.	Rembrandt	Self Portrait	Kunsthist. Museum, Vienna (Hofstede 581)
30.	Rubens	Portrait of a Bearded Man	Liechtenstein, Vienna
31.	Barent Fabritius	Eli and Samuel	Art Institute, Chicago
32.	Cézanne	Seine at Bercy	Kunsthalle, Hamburg (Ven- turi 242)
33.	Rousseau, Henri	Summer	Collection Flachfeld, Paris
34.	van Gogh	Montmartre	Art Institute, Chicago (De la Faille 272)
35.	El Greco	St. Francis and the Skull	Art Institute, Chicago (A. L. Mayer, No. 267)
36.	Corot	The Haywagon	Collection Dollfus (Robaut, No. 1117)
37.	Vermeer	The Lacemaker	Louvre, Paris (Hofstede No. 11)
38.	Manet	Mlle. Victorine as an Espada	Metropolitan Museum, New York
39.	Corot	Morning on the Lake	Robaut, No. 1625
40.	Gauguin	Tahitian Woman with Children	Art Institute, Chicago
41.	Winslow Homer	Adirondacks Guide	Art Institute, Chicago
42.	Carl Hofer	Landscape in the Tessin	
43.	Goya	Boy on a Ram	Art Institute, Chicago
44.	Chardin	Gir Scraping Vege- tables	Liechtenstein, Vienna (Wild- enstein, No. 46)
45.	Vermeer	Lady with a Lute	Metropolitan Museum, New York
46.	Dufy	Regatta at Deauville	Louvre, Paris
47.	Dégas	L'Absinth	Louvre, Paris
48.	Breughel	The Crash of Ikarus	Museum, Brussels
49.	Cézanne	The Aqueduct	Museum of Occidental Art, Moscow (Venturi 477)

TABLE 9

SAMPLE DATA SHEET

School C. High School

Summary for Test

Grade 12Number of Students 20

These are per cent scores				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																
				These are raw scores																

TABLE 10
Summary of the Validity Studies Conducted at Four Schools

Ratings —————>	Number of Students				Description Satisfactory				Description of Questionable Value				Description Unsatisfactory				
	Very good	Partly very good	Partly generally accurate	1-2	1-3	Partly very good	Partly possibly accurate	Generally accurate	Partly very good	Partly generally accurate	Partly of doubtful value	Partly of doubtful value	Partly generally accurate—Partly of doubtful value	Of doubtful value	Partly of doubtful value	Partly very good	Poor
George School.....	17	9						4	1	1			3	1	4	5	5
Cambridge School.....	33	16						2	6								
Granbrook Academy of Art.....	27	18						6									2
Milwaukee State Teachers College.....	31	22						6									
Totals—Individual Categories.....	108	65	4	1	1	18		16	1	1	7	3	1	4	1	2	2
By per cent of students (approximately).....	100	60	4	1	1	16		16	1	1	6	3	1	4	1	2	2
Totals—Grouped Categories.....	108	88								9				11			
By per cent of students (approximately).....	100	81								9				10			

Appendix V

TABLE FOR CHAPTER V

Reliabilities, Means, and Standard Deviations for "Like" of the Different Categories in 8.2a for a Population of 542 Students (261 Boys, 281 Girls) in the 11th Grade

Table 8.2a for a Population of 3,122 Items											
No. of Items in Category	Category		Mean %	Sigma %	r	No. of Items in Category	Category		Mean %	Sigma %	r
24	Soc. Sci.	Total	39.3	27.2	.92	16	Home Econ.	Total	50.8	28.6	.88
		Boys	41.6	27.6				Boys	30.8	21.5	.80
		Girls	37.1	25.4				Girls	69.4	22.2	.81
16	Biology	Total	45.6	27.8	.87	16	Ind. Arts	Total	50.8	25.6	.82
		Boys	44.4	29.4				Boys	59.2	26.2	
		Girls	46.6	26.3				Girls	42.9	23.3	
16	Phys. Sci.	Total	50.1	29.5	.89	16	Fine Arts	Total	45.8	30.0	.89
		Boys	60.4	27.6				Boys	33.2	26.8	.87
		Girls	40.6	28.0				Girls	57.4	28.2	.88
16	English	Total	48.7	26.4	.85	16	Music	Total	46.8	29.4	.89
		Boys	39.5	27.0				Boys	37.0	28.6	
		Girls	57.2	23.0				Girls	55.8	27.3	
16	Foreign Lang.	Total	47.4	31.0	.90	16	Sports	Total	55.2	23.5	.79
		Boys	36.7	29.8				Boys	56.8	24.3	
		Girls	57.2	29.0				Girls	53.6	23.6	
16	Mathe- matics	Total	36.3	29.0	.89	38	Manipu- lative	Total	45.2	19.1	.85
		Boys	45.8	29.2				Boys	41.6	19.2	
		Girls	27.6	25.6				Girls	48.2	18.3	
16	Busi- ness	Total	55.9	23.6	.80	35	Read- ing	Total	47.0	22.6	.90
		Boys	56.8	24.3				Boys	45.8	23.6	
		Girls	55.0	23.9				Girls	48.0	22.5	

Appendix VI

TABLES FOR CHAPTER VI

TABLE 1

Ranges, Means, Standard Deviations, and Reliabilities of the Different Categories in 8.2b from a Random Sample of 1000 Students (7th Grade through 12th Grade)

No. of Items in Category	Category	Likes				Dislikes			
		Range	Mean %	Sigma %	r	Range	Mean %	Sigma %	r
19	Aggression.....	0-94	28.9	18.2	.75	0-99	36.2	19.4	.76
25	Out-of-School Activi- ties.....	0-100	51.8	19.8	.80	0-95	16.9	13.2	.72
32	Family.....	0-100	49.3	19.4	.84	0-90	21.4	14.0	.78
16	Dramatics.....	0-100	34.6	26.2	.86	0-100	27.4	24.2	.86
28	Opposite Sex.....	0-99	48.5	20.2	.85	0-95	19.5	15.5	.79
10	Same Sex.....	0-100	55.0	23.2	.62	0-91	21.9	17.3	.59
18	School Activities.....	0-100	48.8	21.2	.76	0-95	18.9	14.5	.68
26	Authority.....	0-74	21.3	12.1	.62	0-99	47.2	15.8	.71
16	Leadership.....	0-100	35.0	22.8	.79	0-100	22.7	19.8	.79
16	Fantasy.....	0-100	42.2	24.8	.82	0-100	21.1	19.6	.79
16	Maric.....	0-94	26.1	21.4	.80	0-95	28.0	21.8	.80
16	Mystery.....	0-100	39.8	21.8	.77	0-95	23.8	18.4	.75

TABLE 2

Ranges, Means, Standard Deviations, and Reliabilities of the Different Categories in 8.2c from a Random Sample of 1000 Students (7th Grade through 12th Grade)

No. of Items in Category	Category	Likes				Dislikes			
		Range	Mean %	Sigma %	r	Range	Mean %	Sigma %	r
14	Aggression	0-100	31.6	20.8	.73	0-95	32.5	20.2	.71
16	Severity	0-90	33.5	18.6	.70	0-95	29.9	16.6	.64
24	Life-Death-Universe .	0-99	33.0	22.6	.86	0-99	26.9	21.0	.86
26	Preoccupation with Cleanliness	0-90	47.2	18.6	.78	0-80	22.6	13.6	.73
24	Humor	0-99	47.0	19.8	.80	0-80	21.9	15.4	.76
24	Self-acceptance	0-95	42.5	19.4	.78	0-85	28.7	15.8	.72
25	Methodical	0-100	42.0	20.2	.81	0-95	23.0	17.3	.79
16	Identification with Others	0-100	49.0	23.4	.79	0-85	16.9	13.6	.62
16	Non-identification with Others	0-90	34.2	20.0	.73	0-95	30.8	18.0	.66
18	Solitary	0-85	40.0	15.4	.53	0-85	33.5	15.3	.56

INDEX

- Ability, level of, in score interpretation, 435-436
- Ability to Apply Social Facts and Generalizations test, 168-175; behavior, analysis of, in, 172-173; criteria for appraisal, 173-174; objective, analysis of, 168-169
- Achievement, analysis of, in Application of Principles test, 104-111
- Achievement tests, inadequacies of early, 3-4
- Activities: out-of-school, 369; school, evaluation of, 368-369; records, use of, 166; records, reliability and validity of, 330
- Adaptation, role of, in adjustment, 354
- Adjustability (*see also* Adjustment), social, 482
- Adjustment: maturation and adaptation in, 354; meaning of, 350-354; optimum, 353-354
- Administration of Evaluation Program, 439-459
- Administrative problems in obtaining records, 449-450, 454
- Adolescents: interests of, 316; verbal expression of art statements by, 279 n
- Aggression, evaluation of, 366-377
- Aims (*see* Objectives)
- Ambivalence: between general and specific values, 241-242; in social beliefs, 431-432
- Analogy as type of behavior, 49, 101
- Analysis, power and habit of, in Behavior Description, 480
- Analysis of Controversial Writing test, 150-154; conclusion concerning, 154; scoring, 152-154; sample problems in, 151-152; criteria for selecting items, 150
- "Anecdotal method" of recording, 466
- Anecdotal records: criteria for selecting, 163 n; inadequacies of, in evaluating art appreciation, 279; social sensitivity, 160-161, 163-164
- Application of principles of logical reasoning: evaluation instruments, development of, 114-122; objective, analysis of, 111-114 (*see also* Application of Principles of Logical Reasoning test)
- Application of Principles of Logical Reasoning test, 111-126; readiness of class for, 126; sample problem in, 119-121; scores, summary and interpretation of, 122-124; statements, kinds of, in, 121; validity and reliability of, 124-126
- Application of Principles test, 77-111; analogy of statements in, 101; authorities, statements of, in, 101; construction of, 80-111; data sheet sample, 102; directions for, example, 88-89; errors in responses, 83-84; essay-type vs. objective form, 84-85; problem situations in, 80-111; reasons for responses in, 82-84; sample problem, 89-90; scores, summary and interpretation of, 103-111; social values tested in, 95-101; types of responses in, 81-82
- Application of Science Principles (*see* Application of Principles)
- Applying Social Facts and Generalizations to Social Problems test (*see also* Ability to Apply Social Facts and Generalizations), 175, 197-203; behaviors evaluated in, 197-198; description of, 198-203; objective, analysis of, 197-198; sample exercises in, 199-203; uses of, 243

- Appraisal (*see* Evaluation, Reports, Tests)
- Appreciation, Aspects of (*see also* Appreciation of Art, of Literature, of Social Values), 245-312
- Appreciation of art (*see also* Art): evaluation of, 276-312
- Appreciation of literature, 246-276; meaning of, in the Study, 246; behaviors in, 249; test of, 250-276
- Appreciation of social values, use of, 240
- Areas of activity (*see also* Areas of Living), interest tests and, 318
- Areas of Living, interests, role of, in, 317-318
- Areas of thought in Behavior Description, 485
- Argumentum ad Hominem* principle, 112
- Art (*see also* Art Appreciation, Painting, Art Experience, expression, sensitivity), interest in, tests of, 277
- Art Appreciation (*see also* Art Experience): assumptions concerning, 283-285; evaluation of, 276-312; meaning of term, 280-281, 283-284; objectives, 276-277; psychology of, 280-283; records, inadequacies of most, 279; test (*see* Art Appreciation test)
- Art Appreciation test (*see also* Finding Pairs of Pictures test): administration of, 299-300; assumptions underlying, 283-285; criteria for, 279-280; 287-289; description of, 289-292; development of, 283-289; interpretation of, 292-299; reliability of, 300-301; score range, 304; scoring, 292-294; use of, 306-307; validity of, 301-303, 305-306
- Art Experience (*see also* Art): and creativity, 282-283; emotional reaction in, 285; meaning of, 281; methods of data gathering, 277-278; nature of, 285; spectator's role in, 281-283
- Art expression and "Gestalt" psychology, 280-281
- Art History as an Academic Study, 283-284
- Art sensitivity, meaning of, 283-284
- Art teaching, purposes of, 276
- Art test (*see* Art Appreciation test)
- Art values, sensitivity to, 276-277, 278
- Art and verbal facility, 278-279
- Artist's reactions, 283-284
- Arts (*see* Art, Dramatics, Theater)
- Aspects of Appreciation (*see* Appreciation)
- Assumptions, basic, of Evaluation Staff, 11-15
- Assurance in Behavior Description, 485
- Authority, reactions to, 370
- Background data in one case study, 409-410
- Battery of instruments, reasons for, 406-408
- Behavior: central pattern of, 430-431; classifications, 351-352, 484-485; combinations of, determined, 433-434; descriptions (*see* Behavior Description); deviant, hypotheses concerning, 431-432; motivation, role of, in evaluating, 351; objectives defined in terms of, 19-20; organic unity of, 7, 405; patterns, 11-12, 13, 19-20
- Behavior Description, 470-487; advantages of, 486-487; classifications in, 484-485; on college-entrance blank, 496-497; Committee on, 470; data interpretation, functions of, in, 403-404; Manual, 485-486, 496; records, 279, 466, 471, 474-487, 493; in subject fields, 501-502
- Belief: as type of social attitudes, 205; instruments, 208-209
- Beliefs About School Life test, 208, 229-234; results of, in one school, 437
- Beliefs on Economic Issues, characteristics of, 235-236

- Beliefs on Economic Issues test, 208, 234-238
- Beliefs on Housing, 209
- Beliefs on Social Issues test, 208-234; consistency evaluated in, 223; data sheet sample, 221-225; description of, 215-229; evaluation of, 209-215; honesty in, 225-226; language's role in validity of, 225; reliability studies of, 228; sample statements in, 216-217; sampling and statement formulation in, 209-212; score patterns, 224; scores, interpretation of, 220-225; scoring and summarizing results, 217-220; uncertainty evaluated in, 222; validity and reliability of, 225-229
- Beyond data, 55, 56, 62, 408
- Bibliography of evaluation instruments, 21-22 n
- Biology (*see* Science)
- "Birth-Life-Death" fantasies, 370
- Carnegie Foundation for the Advancement of teaching, 494 n
- Carroll, Herbert, 246
- Case study, one, based on test data, 408-429
- Caution score, 55
- Changes (*see also* Growth, Student Growth); behavior, as educational objective, 11-12; diagnoses of, by tests, 242-243; in school practices, resulting from evaluation, 457; in school programs, resulting from evaluation, 436-437; in students, evaluation as check of, 436
- Checklists: data summaries of, in reading interests, 334-337; validity of, 333-334
- Chemistry (*see* Science)
- Classroom situations as source of evaluation data, 446
- Classroom teacher (*see* Teacher)
- Cleanliness, preoccupation with, 366
- "Clear thinking" objectives, 35-37
- College: changes in information for admission to, 495; Committee on Admission, report to, 494-498; "Junior Year" blank for, 498; transfer from school to, form, 494-498
- Committee: on Admission to College, report to, 494-498; on Evaluation in the Arts, 276; on the Evaluation of Interests, 313; on the Evaluation of Interests and Appreciations, 245; on the Evaluation of Reading, 247; on Evaluation and Recording, xx; on the Interpretation of Data, 38; on Reports and Records, 464; on School and College Relations of the Educational Records Bureau, 495; on the Study of Adolescents, 349
- Community (*see also* Home, Parents, Public Relations), evaluation's role in school's relations with, 10
- Compulsiveness, evaluation of, 366
- Conservatism: beliefs, 213; in social beliefs, 217-219; terms, as indicating direction, 213, 217, 220
- Consumer aspect of applying logical principles, 114
- Content, course, as means to ends, 11
- Controversial Writing test, analysis of, 150-154
- Cooperative planning (*see also* Evaluation program planning), 440-442
- Counselor (*see also* Teacher): Interest Questionnaire, value of, to, 345-347, 306-399; interpretation of evaluation data by, 452; interviews with, in one case study, 413-417
- Course revision, evaluation in, 26-27
- Creation (*see also* Creativeness), meaning of, 474
- Creativeness: in art experience, 282-283; in appreciation of literature, 248, 251; characteristics of, 474-475; evaluation of, 475-476; and Imagination in Behavior Description, 474-476, 478
- Critical-mindedness in Reading of Fiction test, 265-267

- "Critical" thinking (*see* Clear thinking, Logical Reasoning)
- "Crude errors" in data interpretation, 47, 55
- Cultural activities in one case study, 417-418
- Curiosity in Appreciation of Literature test, 248, 251
- Curriculum: based on hypotheses, 7-8; changes in, resulting from evaluation, 436-437; effectiveness of, appraised, 453-454; improvement of, one purpose of evaluation, 403, 432-436; Reading questionnaires in appraising, 275-276; and school program (*see* School program)
- Dale, Edgar, 328 n
- "Dartmouth Visual Survey," 473 n
- Data (*see also* Evaluation Data); classifications of, 41-42; criteria for selection, 42-43; dependability of, evaluating, 40; interpretation of, 38-76; kinds of, for interpretation, 41-43; presentation of, forms, 41; selection and use of, 31-32; sources of, 42
- Deductive thinking, 78
- Definitions principle, 112
- Democracy (*see also* Democratic): as interest area in social issues, 209; liberalism and conservatism regarding, 217; in school, 229
- Democratic: meaning of term, 183; attitudes, evaluation data useful in developing, 457; tenets (*see also* Social Problem values), 175, 179; values appraised in Social Problems test, 183-184, 187
- Descriptive Trait Profile, 358, 383-384, 388
- Devices (*see* Instruments, Tests)
- Directing Committee of the Study, 3-4
- Drama Questionnaire, 253, 264
- Dramatics, interest in, 371-372
- Drives and impulses, organization of, 364-367
- Economic issues (*see also* Beliefs on Economic Issues test), beliefs on, 234-238
- Economic relations: as interest area in social issues, 209; liberalism and conservatism regarding, 217-218
- Education: continuity of, 494-495; purpose of, 11
- Eells, Walter Crosby, 326
- Emotional adjustment fostered by the arts, 276
- Emotional control in Behavior Description, 485
- Emotional disposition and "academic" interests, 396
- Emotional Responsiveness, in Behavior Description, 481
- Emotional tendencies, interpretation of (*see* Interests and Activities Questionnaire, interpretation of) "Empathy," 280
- Environment and individual, relationship, 468
- Essay-type test: criticisms of, 84-85; and Form 2.52, correlation between, 67-73
- Esthetic experience, 280, 283
- Evaluating, habit of, 33
- Evaluation (*see also* Evaluation Data Tests): continuity of, essential, 438, 442; complexities, reasons for, 6-7; definition of, in the Study, 5; influences of, on teaching and learning, 14; interpretation of (*see also* Interpretation), 6, 25-28; methods, selection of, 21-23; role of, in educational process, 29-30; purposes of, 7-11, 403, 432-437; results of, use of, 454-459; school's responsibility for, 14; traditional, inadequacies of, 146; whole-faculty responsibility for, 438
- Evaluation adviser, 458
- Evaluation of Art Appreciation (*see also* Art Appreciation), 276-312
- Evaluation data: assumptions underlying, 405-408; available in planning program, 445; case study

- illustrating synthesized, 408-429; circulation of, 444-447, 449-454; collection of, methods, 444; faculty attitude toward, 445-456; in guidance, 430-432; interpretation of, and teachers, 437-438; interpretation and user of, 403-438; nature of, 405-408; sources of, 446; summarizing and circulating, 449-454; synthesized, case study illustrating, 408-429; uses and interpretation of, 403-438
- Evaluation devices (*see also* Evaluation instruments, Tests), developing and improving, 23-25
- Evaluation Instruments (*see also* Tests): Bibliographies of, 21-22; development of, 43-60; need for new, 4
- Evaluation of Interests (*see also* Interests), 313-348
- Evaluation of Personal and Social Adjustment (*see also* Personal and Social Adjustment test), 349-402
- Evaluation Program: concept of, by teachers, 442; division of labor in, 28-29; interpretation and uses of data in (*see also* Data, Interpretation), 403-438; as integral part of school, 459; limitations in planning, 443-444; as method of teacher education, 30; misconceptions about, 442; needs served by, 443; planning and administering, 439-459; procedures in developing, in the Study, 15-28; purpose of, 442; scope and emphasis of, 441-444; summary of, 459; summary of planning and administering, 439
- Evaluation specialist, inadvisability of having, 440
- Evaluation Staff: basic assumptions by, 11-15; members of, 4, 5
- Evaluation techniques, wide range of, needed, 13-14
- Examinations (*see* Tests)
- Experimentation in creativeness, 475
- Extrapolation, 39, 45-46
- Faculty (*see also* Counselor, School Staff, Teacher): attitude of, toward evaluation data, 455-456; continuity of study and collective thinking by, 454-456; participation of whole, in evaluation program planning, 441, 457-458; responsibility of whole, in securing data, 446
- Family relationships, evaluation of, 367
- Fantasy: "Birth-Life-Death," 370; behavior, 370-371; in Interest and Activities Questionnaire, 364, 370-372
- Feelingtone, type of social attitude, 205
- General accuracy, definition of, in test response, 51
- General science (*see also* Application of Principles test, Science), test construction for applying principles in, 80-111
- Generalizations (*see also* Application of Social Facts and Generalizations), testing for formulation of, 24-25
- "Gestalt" psychology and art expression, 280-281
- Grades and awards (*see also* Marks, Reports), as area in Beliefs about School Life Test, 232
- Group: life, as area in Beliefs about School Life, 231; progress, processes to estimate, 433-436
- Growth (*see also* Changes, Pupil growth); group's, measure of, 433-436; individual, reports of, 489-490
- Guidance (*see also* Counselor) continuity of fostered by records, 465; evaluation data, use of, in, 8-9, 430-432, 454-455; reports in, 492; and transfer, recording for, 463-504
- Gullibility, 153
- Habits, work, appraisal methods needed for, 31-33

- Home reports (*see also* Parents, Reports), 488-493; records as bases of, 465; and teacher reports, identical, 492
- Homeroom teacher, evaluation data summaries interpreted by, 452
- Hoskins, Luella, 329 n
- Housing, Beliefs on, test, 209
- Human Behavior (*see* Behavior)
- "Human Relationships" as area in Interests and Activities Questionnaire, 364, 367-370
- Humor, activities in expressions of, 372
- Hypotheses, validation of, as one purpose of evaluation, 7-8
- Identification: in appreciation of literature, 248, 251; with others, evaluation of, 368
- If-then principle, 112
- Imagination in creativeness, 475
- Impressing others, activities in, 369
- Impulses and drives, organization of, as area in Interests and Activities Questionnaire, 364-367
- Indirect argument principle, 112
- Inferences: in data interpretation, 39; test to measure, 60-62
- Influence in Behavior Description, 478-479
- Inquiring Mind in Behavior Description, 479
- "Insight," meaning of term, 398-399
- Instruments (*see* Evaluation Instruments, Tests)
- Intelligence, general, relation of, to Social Problems test results, 196
- Intercorrelation of scores in Interpretation of Data test, 59 n
- Interest (*see also* Interest Index, Interest Questionnaire, Interests): and appreciation, distinctions between, 245; in Reading test, validity and reliability of, 330-334
- Interest Index (*see also* Interest Questionnaire, Interests and Activities Questionnaire), 338-348; areas in, 339; in one case study, 415; data sheet sample, 341; interpretation of, 340-345; uses of, 347-348
- Interest Questionnaire (*see also* Interest Index, Interests and Activities): analysis of, sample, 377; and checklists, 337; construction of, 338-340; use of, in developing personality test, 358-359, 360-361; value of, to counselor and teacher, 345-347
- Interests (*see also* Interest, Interest Index, Interest Questionnaire, Interests and Activities, Recreational Interests): "academic", and emotional dispositions, 396; adolescent vs. adult, 316; data sources for revealing, 313; evaluation of, 313-348; as index of personality pattern, 359; as means and ends, 313-314; objectives, analysis of, 313-318; questionnaire (*see also* Interest Questionnaire), 338-348; patterns of, as revealed by checklists, 334-337; recreational (*see* Recreational Interests); significance of, in personality evaluation, 359-360; uniqueness of, 344
- Interests and Activities Questionnaire (*see also* Interest Questionnaire): administration of, 400; areas in, 364; categories in, 363-372; criteria for item selection in, 362-363; drives and impulses, organization of, as area in, 364-367; interpretation of, 372-384, 390-392; interpretation of, to students, 399; validity of, 387-396
- Interpolation in interpreting data, 39
- Interpreter, importance of, in tests, 154-155
- Interpretation (*see also* Interpretation of Data): ability to make original, 67-74; ability to judge by others, 65-67; behavior descriptions, one function of, 403-404; functions of, 403-405; over-all, by

- staff member, 452-453; overgeneralized, 46; undergeneralized, 47
- Interpretation of Data (*see also* Interpretation of Data test), 38-76; accurate, 46; classifications of types of, 46-47; original vs. stated, 40-41, 65; types of, 45-46
- Interpretation of Data test, 47-60; appropriateness of, for high-school level, 67; construction of, 48-51, 66; form of, for junior high school, 63-65; forms of, 67-73; reliability of, 74-76; response patterns to, 73; validity of, 65-76
- Interpretation and Uses of Evaluation Data, 403-438
- Interschool Committee, 28-29
- Judging the Effectiveness of Written Composition test, 265, 267-268; Junior High school: Application of Principles test for, 91; n; Interpretation of Data test for, 63-65
- "Junior Year" blank, 498
- Kuder-Richardson formula, 65
- Labor and unemployment: as interest area in social issues, 209; liberalism and conservatism regarding, 218
- Language (*see also* Words): choice of, in statements of social beliefs, 210-212
- Leadership, activities in, 369
- Learning, influence of evaluation on, 14
- Liberalism: meaning of term, 213, 217, 220; in social beliefs, 217-219
- Life, philosophy of, appraised, 34
- Literature (*see also* Appreciation of Literature, Recreational Interests), appreciation of, 246-276
- Logical reasoning: behaviors in, 112-113; meaning of, 111-112; test of (*see* Application of Principles of Logical Reasoning)
- Magazines (*see also* Reading magazines): checklist of, 326; classification of, by types, 326
- Maladjustment (*see also* Adjustment, Personal and Social Adjustment), kinds of, 353
- Manipulation in creativeness, 475
- Marks (*see also* Grades and Awards, Home Reports, Parent Reports, Reports, Teacher Reports): for college admission, inadequacies of, 488-489, 494; and interests, 316; as objectives, 494; in records and reports, 467, 468
- Maturation, role of, in adjustment, 354
- Methodical activities, evaluation of, 366
- Methods, evaluation: means to ends, 11; selection and trial of, 21-23
- Militarism: as interest area in social issues, 209; liberalism and conservatism regarding, 218
- Motivation: personal and social adjustment study yields insight into, 401; role of, in evaluating behavior, 351
- Movies, checklists for revealing recreational interests regarding, 328-329
- Mystery-interests, 371
- Nationalism: as interest area in social issue, 209; liberalism and conservatism regarding, 219
- Nature of proof, 36; assumptions in, 129; behaviors in achieving, 129; definition of, 127-129; objective, analysis of, 126-130; senses in arriving at, 128; test (*see* Nature of Proof test)
- Nature of Proof test (*see also* Nature of Proof), 126-148; check on responses to, 144-147; development of, 130-141; sample problems in, 132-134, 136-139; scores, summary and interpretation of, 141-143; structure of, 135-141; validity and reliability of, 143-148

- Objectives: agreement on, needed for evaluation, 6; analysis of, 38-43; application of logical reasoning, analysis of, 111-114; application of scientific principles, 77, 111; an evaluation program, areas of, 406; "breaking up," 405; changes in behavior patterns as, 11-12; classification of, 16-18; "clear thinking" as, 35-37; "comprehensive," 406; concern in evaluation, 5; defining, in terms of behavior, 19-20; evaluation data collection regarding, 444-446; evaluation program as a check on achievement of, 432-436; formulation of, 15-16; general and specific, relation between, 441-442; of growth reports, 489-490; illustrations of, 12; "intangible," 439; interests as, 317-318; interests and appreciations as, 245; limited overemphasis on, reasons for, xvi; marks as, 494; propaganda analysis, 149-150; record forms for, in subject fields, 500-502; in records, 463-469; re-examination of, essential, 16; reformulation of, 30; selection of, basis for, 15-16; situations showing achievements of, 20-21; statements of, inadequacies of, xv; in subject fields, study of, 499-500; teacher consideration of, 465-466; types of, 18; working, for records and reports, 467-469
- Omissions, scoring of, in test, 55
- Open-mindedness in Behavior Description, 479-480
- Organization of Impulses and Drives, as area in Interests and Activities Questionnaire, 364-367
- Out-of-school activities, evaluation of, 369
- "Overcaution" in data interpretation, 47
- "Overcritical" students, 97
- Painting (*see also* Appreciation of Art, Art), field of, chosen for art test, 286
- Parents (*see also* Community, Home Reports): participation of, in suggesting areas of social beliefs, 207; reading questionnaires, results of, for parents, 275; reports to, 488-493; reports of evaluation useful to, 456; security of, fostered by evaluation, 9-10
- Pencil-and-paper tests, 44; use of, in collecting evaluation data, 446-447
- Personal adjustment (*see also* Adjustment, Maladjustment, Personality): meaning of, 350; and social adjustment (*see* Personal and Social Adjustment)
- Personal and social adjustment (*see also* Interests), 206; appraisal, techniques of, 354-358; cleanliness, preoccupation with, in, 366; differentiation between, 350; evaluation of (*see also* Personal and Social Adjustment test), 349-402; interests, significance of, in, 359-360; objective, history of, 349-350; summary regarding, 400-402
- Personal and Social Adjustment test: characteristics, desirable, of, 355-358; Interest Questionnaire, use of, in developing, 358-359, 360-361; Interests and Activities Questionnaires for, 361-402
- Personality (*see also* Adjustment, Personal Adjustment, Personal and Social Adjustment): one case study of, 376-384; information about, need for, xix; meaning of term, 350; measurement of, 351 n; projective methods for studying, 36; rating scale, 358
- Philosophy and objectives underlying recording, 463-469
- Philosophy of life, appraisal of, 34
- Physical energy in Behavior Description, 485
- Physics (*see* Science)
- Planning and administering the evaluation program, 439-459
- Prejudices in social attitudes, 206

- Pre-tests, 238, 243
- Primitive drives and impulses, 364-365
- Principles of logical reasoning (*see* Application of Principles of Logical Reasoning tests, Logical reasoning)
- Programs (*see also* Evaluation program), evaluation data useful in making out, 456
- Proof, nature of (*see* Nature of proof)
- Propaganda: definitions of, 149; analysis, 148-154; behaviors related to, 149-150
- Public relations (*see also* Community, Home, Parents), evaluation as a basis for, 10
- Pupil (*see also* Student): development in subject fields, 499-504; growth, objectives of, classification of reports on, 490; and teacher relations, 231
- Purposes (*see* Objectives)
- Qualitative vs. quantitative understanding, 82
- Questionnaire; techniques, assumptions in, 252; on voluntary reading (*see* Questionnaire on Voluntary Reading)
- Questionnaire on Voluntary Reading, 253-264; criteria for item selection on, 255-257; data sheet, sample, 259; description of, 253-257; scoring, 258-264; summarizing, 257-264; use of, 273-275
- Race: as interest area in social issues, 209; liberalism and conservatism regarding, 218
- Radio: checklists for revealing recreational interests, 328, 329-330; preferences, 329-330
- "Rating" (*see also* Grades and awards, Marks), 486
- Reading (*see also* Appreciation of Literature, Reading Record, Recreational Interests): fiction, classification of, by type, 322, 324; checklist of, interests, 334-337; magazines, 325-327; non-fiction, classification of, 322, 325; points, 45; reactions to (*see* Reading reactions, Reading Reactions Questionnaire), records (*see* Reading records); voluntary (*see* Reading Questionnaire)
- Reading reactions (*see also* Reading Reactions Questionnaire): evaluation, need for, 249, 252; meaning of, 248-249; synthesis of data in one case study, 417-418; tests of, 265; types of, 248-249, 251-252
- Reading Reactions Questionnaire, 250-273; "direct" forms of, 271-272; direct observations and questionnaire techniques, difference between, 269-270; student honesty in, 269-270; uses of, 273-276; validity of, 268-273
- Reading records: for revealing interests, 319-325; samples of classification in, 322; use of, 166
- Reasoning, logical (*see* Application of principles of logical reasoning, Logical reasoning)
- Record forms: objectives for, 467-469; for objectives in subject fields, 500-501; purpose of, 464
- Record keeping, decentralized, inadequacies of, 450-451
- Records (*see also* Behavior Description, Record forms): activities, 166; activity, validity and reliability of, 330; behavior, 206 n; observational, 43, 449-450; reading (*see also* Reading records), 319-325; and reports, objectives, 467-469
- Recreational interests: areas of, 318; checklists, use of, 334-337; magazine checklist for revealing, 325-326; movie checklists for revealing, 328-329; newspaper questionnaire for revealing, 327-328; radio checklist for revealing, 328-330; reading record for revealing, 319-325; validity and reliability of tests for, 330-334

- Reflective thinking (*see also* Application of Principles of Logical Reasoning), process of, 36
- Relationships: family, evaluation of, 367; human, as area in Interests and Activities Questionnaire, 364, 367-370; with opposite sex, evaluation of, 367; with same sex, evaluation of, 367; in social values, 175
- Reliability of scores, 63
- Religion, beliefs on, 208 n
- Report cards (*see* Report forms)
- Report forms (*see also* Reports): 490-491; traditional, 488
- Reports: objectives of, 489-490; to parents, 456, 480-493; on pupil growth, classifications of objectives in, 490; records as basis and part of, 405
- Reports and Records: Committee on, 464; objectives for, 467-469
- Responsibility-Dependability in Behavior Description, 477-478
- Sampling, as type of data interpretation, 46
- Satisfaction, evaluation of, in Appreciation of Literature test, 248, 251
- Scales of Beliefs (*see also* Beliefs): 207, 239; on economic issues (*see* Beliefs on Economic Issues test); on social issues (*see* Beliefs on Social Issues test); uses of, 240-241
- Schedule for testing, 447-448
- School: democracy in, 229; evaluation, responsibility of, for, 14; evaluation of activities in, 368-369; government as area in Beliefs about School Life, 230-231; life, beliefs about (*see also* Beliefs about School Life), 208; objectives (*see* Objectives, school); program (*see also* Curriculum), changes in, 436-437; program, hypotheses underlying, 436; resources, evaluation program limited by, 443-444; spirit, 233; staff (*see also* Faculty, Teachers): security of, fostered by evaluation, 9-10; training of, for interpreting evaluation results, 27-28
- Science principles (*see also* Application of Principles): application of, 77-111; meaning of, 78-80
- Score (*see also* Scores): "beyond data," 55; caution, 55; crude errors, 55; omissions in, 55; derivation of, 54; general accuracy, 51
- Scores: analysis of, on Interpretation of Data test, 56-60; intercorrelation of, 59 n; reliability of, increased, 63; students' knowledge of, inadvisable, in Interests and Activities Questionnaire, 397-399; summary and interpretation of, in Application of Principles test, 103-111
- Secondary school (*see* School)
- Security, psychological, fostered by evaluation, 9-10
- Self-reliance in Behavior Description, 485
- Senses, use of, in arriving at proofs, 128
- Seven Modern Paintings test, 307-312
- Short-answer tests, 47
- Social action, skill in securing evidence of, 161-168
- Social adjustability: in Behavior Description, 482; meaning of, 350
- Social attitudes: analysis of behavior in, 204-209; belief, as type of, 205; characteristics of, 205; definition and classification of, 161, 203-209; evaluation of (*see* Applying Social Facts and Generalizations); expressions of, 206-207; feeling-tone as type of, 205; tendency to act, as type of, 205
- Social awareness, meaning of, 161
- Social beliefs (*see also* Beliefs on Social Issues): ambivalence in, possible reasons for, 431-432; areas of, 207-209; characteristics of, 212-214; conservatism in, 217-

- 219; consistency of, 213-214; instruments of evaluation, 208-209; liberalism in, 217-219; scale construction of, 214-215; about school life (*see* Beliefs about School Life); statements of, language in, 210-212; test (*see* Beliefs on Social Issues, Social sensitivity); threshold in statement of, 210
- Social consciousness (*see* Social sensitivity)
- Social generalizations, 169-172
- Social information, meaning of, 161
- Social issues (*see also* Beliefs on Social Issues): areas of, 208; areas of interest in, 209-210; direction of positions toward, 212-213; test (*see* Beliefs on Social Issues test)
- Social Problems test (*see also* Beliefs on Social Issues test): comprehensiveness appraised, 174, 184-186; consistency appraisal in, 174, 184, 189-190; criteria for choosing items in, 176; data sheet sample, 185; democratic values appraised in, 183-184, 187; development of, 177-184; intelligence, relation of, to results of, 196; key for, 182-183; logical aspects, interpretation of, in, 189; rationalization appraised in, 188; relevance appraisal in, 174, 187; results of, related to interests, 318; results, summarized, 184-190; scoring, validity of, 191-192; structure for, 176; student interviews, as validity checks of, 194-195; teachers observations compared with results of, 193-194; use of, 240, 241, 244; validity of construction of, 191; validity and reliability of, 190-197; value patterns in, 177, 179, 183, 189
- Social science, generalizations taught in, list of, 169-170
- Social sensitivity: anecdotal records in obtaining evidence of, 160-161, 163-164; aspects of, 159-162; behaviors involved in, 158, 159-162; evaluation of, 157-244; free-response tests in obtaining evidence for, 166; meanings of, 158-159; objectives, origin and scope of, 157-159; pattern of, 166-167; students' writings as means of securing evidence about, 164-166
- Social values (*see also* Beliefs on Social Issues test), 158-159; application of, 175-197, 406; application of, test construction on, 175-180; behavior in applying, 174; beliefs test, uses of, 238-244; tests (*see* Application of Social Facts and Generalizations, Social Problems); use of tests, 238-244
- Society, demands of, conforming to, 352-353
- Solitary activities, evaluation of, 369
- Strong Vocational Interest Blank, 318
- Student (*see also* Pupil) Background of, important in social tests, 233; behavior patterns, organization, 12-13; development, evidence of, sources for, 444-449; interviews, as validity checks, 194-195, 331-333; participation of, in test construction, 207, 211; philosophy of life, appraisal of, 34; programs, evaluation useful in making out, 456; scores on Interests and Activities Questionnaires, unwise to show to, 397-399; security fostered by evaluation, 9-10; self-observations in test construction, 252
- Study, conditions for effective, 31; skills and work habits needing appraisal, 31-33
- Subject fields: objectives in, recording of, 499-500; record forms for, 500-501, 503, 504
- Suggestibility, 152
- Teacher (*see also* Counselor, Faculty Teachers, Teaching): education through evaluation programs, 30; and pupils, relation, 231-232; rat-

- ing and test results compared, 193-194; reports, 410-413, 492; training, 459
- Teachers: concern of, in behavior development, essential, 238, 239; evaluation program, reaction, 20, 432; insights translated into practice by, 454-455; Interest Questionnaire, value of, to, 345-347; as interpreters of evaluation data, 458-459; objectives considered by, 465-466; observations of, compared with test results, 193-194; realization of objectives in subject fields by, 499-500; security of, fostered by evaluation, 9-10; subject-field forms useful to, 504; training of, in interpreting evaluation results, 27-28
- Teaching (*see also* Guidance): evaluation data used in, 454-455; influence of evaluation on, 14
- Tension, inadvisable to point out, to student, 398
- Test (*see also* Evaluation, Tests): construction, 114-122; data, summary of, in one case study, 415-429; responses, terminology describing, 51; scoring, traditional inadequacies of, 44; schedule for, 447-448; situation, total, 156
- Tests: achievement, inadequacies of early, 3-4; allocation of, to faculty, 448; bibliographies of, 21-22; essay-type, criticisms of, 84-85; pencil-and-paper, 44, 162, 187; science, principles of applying, 77-111; readministration of, 155, 156; short-answer, 47; structure of, interpreter's understanding of, 154-155; written, shortcomings of, 14-15
- Theater arts (*see also* Dramatics), interest in, 371-373
- Thinking (*see also* Application of Principles of Logical Reasoning, Clear thinking, Logical reasoning, Social thinking): Aspects of, 35-156; as objective, 35
- Thirty Schools (*see also* School), responsibility of, for evaluation, 3, 14-15
- Thurstone, L. L., 214
- Time: effective use of for study, 31; recording data, economy of, in, 45, 454
- Traits (*see also* Behavior, Behavior Description), 470, 473
- Transfer: Behavior Description card useful in, 486-487; to college, form, 494-498; recording for, 465, 494-498
- Trends, recognition and comparison of, 45-46, 49
- "Undemocratic" (*see also* Democracy, Democratic), meaning of term, 183
- Units for college admission, inadequacies of, 494
- Value: judgment, 45; pattern, ambivalence in, 244
- Values (*see also* Social Values), general vs. specific, 241-242
- Verbal facility and art, 278-279
- Vocabulary, appropriateness of, in administering tests, 239
- Vocational tests, interests sampled by, 318
- Work habits: in Behavior Description, 482; and study skills needing appraisal, 31-33
- Wert, James E., 327
- Whole-faculty (*see* Faculty)
- Wickman, E. K., 353 n
- Words, "people-describing," 470-471

Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological
Research Library.**

The book is to be returned within the date stamped last.

- 5 JUN 1969

WBGP-59/60-5119C-5M

371.25

Form No. 4

BOOK CARD

Coll. No. 371.1.25/SMI Accn. No. 839

Author. Smith

Title. A. P. Smith & Co. Albany
Sellers & Co. Albany

Date.	Issued to	Returned on
- 5 JUN 1961	6261	- 5 JUN 1961
.....
.....

~~370.97~~ 371.25
SMI